

# **Evidence-based Software Engineering using R**

**Derek M. Jones**  
[derek@knosof.co.uk](mailto:derek@knosof.co.uk)



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Software ecosystems . . . . .	2
1.1.1	What are the software engineering activities? . . . . .	4
1.2	Brief history of software engineering research . . . . .	5
1.2.1	The academic ecosystem . . . . .	6
1.2.2	Paper citation practices . . . . .	8
1.3	Lessons learned from the analysis in this book . . . . .	8
1.4	Overview of contents . . . . .	9
1.4.1	Why use R? . . . . .	11
1.5	Terminology, concepts and notation . . . . .	12
<b>2</b>	<b>Human cognitive characteristics</b>	<b>15</b>
2.1	Introduction . . . . .	15
2.1.1	Models of human cognitive performance . . . . .	17
2.1.2	Embodied cognition . . . . .	17
2.2	Motivation . . . . .	18
2.3	Memory systems . . . . .	19
2.3.1	Short term memory . . . . .	20
2.3.2	Episodic memory . . . . .	23
2.3.3	Forgetting . . . . .	23
2.3.4	Recognition and recall of information . . . . .	23
2.3.4.1	Serial order information . . . . .	24
2.4	Learning and experience . . . . .	24
2.4.1	Belief . . . . .	26
2.4.2	Category knowledge . . . . .	27
2.4.2.1	Categorization consistency . . . . .	29
2.4.3	Expertise . . . . .	30
2.5	Visual processing . . . . .	32
2.6	Reasoning . . . . .	35
2.6.1	Deductive reasoning . . . . .	37
2.6.2	Linear reasoning . . . . .	38
2.6.3	Causal reasoning . . . . .	39
2.7	Number processing . . . . .	39
2.7.1	Problem size and symbolic distance effect . . . . .	41

2.8	Human factors . . . . .	41
2.8.1	People make mistakes . . . . .	41
2.8.2	Cognitive effort . . . . .	42
2.8.3	Personality & intelligence . . . . .	43
2.8.4	Attention . . . . .	43
2.8.5	Risk taking . . . . .	44
2.8.6	Decision-making . . . . .	44
2.8.6.1	Expected utility and Prospect theory . . . . .	47
2.8.6.2	Overconfidence . . . . .	48
2.8.6.3	Time discounting . . . . .	49
2.8.7	Personal integrity . . . . .	49
2.8.8	Miscellaneous characteristics . . . . .	49
2.9	Developer performance . . . . .	50
2.10	Biased behavior . . . . .	51
<b>3</b>	<b>Cognitive capitalism</b>	<b>53</b>
3.1	Introduction . . . . .	53
3.1.1	Some definitions . . . . .	53
3.2	Investment decisions . . . . .	54
3.2.1	Discounting for time . . . . .	54
3.2.2	Taking risk into account . . . . .	55
3.2.3	Incremental investments and benefits . . . . .	56
3.2.4	Quality adjusted process . . . . .	57
3.2.5	Information asymmetry . . . . .	57
3.3	Company economics . . . . .	58
3.3.1	Cost accounting . . . . .	59
3.3.2	The shape of money . . . . .	59
3.3.3	Valuing software . . . . .	60
3.4	Maximizing profit . . . . .	60
3.4.1	Product/service pricing . . . . .	61
3.4.2	Predicting sales volume . . . . .	62
3.4.3	Managing customers as investments . . . . .	64
3.4.4	Commons-based peer-production . . . . .	64
3.5	Game theory . . . . .	64
<b>4</b>	<b>Ecosystems</b>	<b>67</b>
4.1	Introduction . . . . .	67
4.1.1	Evolution . . . . .	68
4.1.2	Software ecosystems . . . . .	70
4.1.2.1	Countries . . . . .	70
4.1.3	Evolving data about evolving systems . . . . .	71
4.2	Culture . . . . .	71
4.2.1	Software culture . . . . .	72
4.2.2	Folklore . . . . .	73

4.2.3	Expertise . . . . .	74
4.2.4	Organizational learning and forgetting . . . . .	76
4.3	Customer ecosystems . . . . .	76
4.3.1	Hardware ecosystems . . . . .	78
4.4	Vendor ecosystems . . . . .	79
4.4.1	Companies . . . . .	79
4.4.2	Cooperative competition . . . . .	80
4.5	Career ecosystems . . . . .	80
4.5.1	Career progression . . . . .	82
4.6	Product ecosystems . . . . .	83
4.6.1	Product customization . . . . .	84
4.6.2	Maintenance . . . . .	84
4.6.3	Forking . . . . .	86
4.6.4	Product obsolescence . . . . .	86
4.6.5	Documentation . . . . .	87
4.7	Software development ecosystems . . . . .	88
4.7.1	Programming languages . . . . .	89
4.7.2	Libraries and packages . . . . .	91
4.7.3	Licensing . . . . .	92
4.8	Evolution of source code . . . . .	93
4.8.1	Source code lifetime . . . . .	94
4.8.2	Refactoring . . . . .	95
4.8.3	Software reuse . . . . .	95
4.8.4	Database schema . . . . .	96
4.9	Population dynamics . . . . .	97
4.9.1	Estimating population size . . . . .	98
<b>5</b>	<b>Projects</b>	<b>101</b>
5.1	Introduction . . . . .	101
5.1.1	Project pecking-order . . . . .	103
5.1.2	Cancellation . . . . .	104
5.2	Resource estimation . . . . .	104
5.2.1	Estimation models . . . . .	106
5.2.2	Money . . . . .	108
5.2.3	Time . . . . .	108
5.2.4	Size . . . . .	108
5.3	Pitching for projects . . . . .	109
5.3.1	Contracts . . . . .	110
5.4	The path to delivery . . . . .	111
5.4.1	Development methodologies . . . . .	113
5.4.2	Managing progress . . . . .	114
5.4.3	Major activities . . . . .	116
5.4.4	Discovering functionality needed for acceptance . . . . .	117

5.4.5	Implementation . . . . .	118
5.4.6	Deployment . . . . .	120
5.4.7	Maintenance . . . . .	120
5.5	Development teams . . . . .	121
5.5.1	New staff . . . . .	122
<b>6</b>	<b>Reliability</b>	<b>123</b>
6.1	Introduction . . . . .	123
6.1.1	It's not a fault, it's a feature . . . . .	125
6.1.2	Why do faults occur? . . . . .	125
6.1.3	Reported fault data . . . . .	126
6.1.4	Cultural outlook . . . . .	127
6.2	The search for profit . . . . .	128
6.3	Experiencing a fault . . . . .	131
6.3.1	Input profile . . . . .	131
6.3.2	Further fault experiences—closed population . . . . .	133
6.3.3	Further fault experiences—open population . . . . .	134
6.4	Where is the mistake? . . . . .	136
6.4.1	Human variability—the random walk of life . . . . .	137
6.4.2	Requirements . . . . .	138
6.4.3	Source code . . . . .	139
6.4.4	Third-party libraries . . . . .	141
6.4.5	Correctness vs. consistency . . . . .	141
6.4.6	Documentation . . . . .	141
6.5	Non-software causes of unreliability . . . . .	141
6.5.1	System availability . . . . .	143
6.6	Checking for intended behavior . . . . .	143
6.6.1	Review meetings . . . . .	144
6.6.2	Testing . . . . .	145
6.6.2.1	Combinatorial testing . . . . .	146
6.6.2.2	Beta testing . . . . .	146
6.6.2.3	Estimating test effectiveness . . . . .	147
6.6.3	Cost of testing . . . . .	148
6.6.4	Runtime issues . . . . .	148
<b>7</b>	<b>Source code</b>	<b>149</b>
7.1	Introduction . . . . .	149
7.1.1	Quantity of source . . . . .	151
7.2	Desirable characteristics . . . . .	152
7.2.1	Programming languages . . . . .	152
7.2.2	Understanding code . . . . .	154
7.2.3	Memory for material read . . . . .	155
7.2.4	Integrating information . . . . .	157
7.2.5	Visual organization . . . . .	159

7.2.6 Identifier names . . . . .	160
7.2.7 Consistency . . . . .	163
7.2.8 Folklore metrics . . . . .	164
7.3 Patterns of usage . . . . .	164
7.3.1 Language features . . . . .	166
7.3.2 Runtime characteristics . . . . .	167
7.3.3 Narrative structures . . . . .	168
7.3.4 Control flow . . . . .	168
7.3.5 Statements . . . . .	169
7.3.5.1 Literal values . . . . .	169
7.3.5.2 Calls . . . . .	170
7.3.5.3 Use of variables . . . . .	170
7.3.6 Declarations . . . . .	171
7.3.6.1 Visibility . . . . .	171
7.3.6.2 Ordering of definitions in aggregate types . . . . .	171
7.3.6.3 Unused variables . . . . .	172
7.4 Building programs . . . . .	172
<b>8 Stories told by data</b>	<b>173</b>
8.1 Introduction . . . . .	173
8.2 Finding stories in data . . . . .	174
8.2.1 Initial data exploration . . . . .	174
8.2.2 Guiding the eye through the data . . . . .	179
8.2.3 Smoothing data . . . . .	180
8.2.4 Densely populated measurement points . . . . .	181
8.2.5 Visualizing the distribution of values . . . . .	183
8.2.6 Relationships between items . . . . .	184
8.2.7 3-dimensions . . . . .	185
8.3 Communicating a story . . . . .	188
8.3.1 What kind of story? . . . . .	190
8.4 Technicalities should go unnoticed . . . . .	192
8.4.1 People have color vision . . . . .	192
8.4.2 Color palette selection . . . . .	193
8.4.3 Plot axis: what and how . . . . .	194
8.5 Communicating numeric values . . . . .	195
8.5.1 Percentages vs frequencies . . . . .	196
<b>9 Probability</b>	<b>197</b>
9.1 Introduction . . . . .	197
9.1.1 Useful rules of thumb . . . . .	198
9.1.2 Measurement scales . . . . .	199
9.2 Probability distributions . . . . .	200
9.2.1 Comparing probability distributions for equality . . . . .	204
9.3 Fitting a probability distribution to a sample . . . . .	206

9.3.1	Zero-truncated and zero-inflated distributions . . . . .	208
9.3.2	Mixtures of distributions . . . . .	209
9.3.3	Heavy/Fat tails . . . . .	211
9.4	Markov chains . . . . .	211
9.4.1	A Markov chain example . . . . .	212
9.5	Social network analysis . . . . .	213
9.6	Combinatorics . . . . .	214
9.6.1	A combinatorial example . . . . .	215
9.6.2	Generating functions . . . . .	216
<b>10</b>	<b>Statistics for software engineering</b>	<b>219</b>
10.1	Introduction . . . . .	219
10.1.1	Statistical inference . . . . .	220
10.2	Samples and populations . . . . .	221
10.2.1	Sampling error . . . . .	222
10.3	Describing a sample . . . . .	223
10.3.1	A central location . . . . .	223
10.3.2	Sensitivity of central location algorithms . . . . .	225
10.3.3	Geometric mean . . . . .	225
10.3.4	Harmonic mean . . . . .	226
10.3.5	Contaminated distributions . . . . .	226
10.4	Statistical error . . . . .	227
10.4.1	Hypothesis testing . . . . .	227
10.4.2	p-value . . . . .	229
10.4.3	Confidence intervals . . . . .	229
10.4.4	The bootstrap . . . . .	231
10.4.5	Permutation tests . . . . .	231
10.5	Effect-size . . . . .	232
10.6	Statistical power . . . . .	233
10.7	Meta-Analysis . . . . .	235
<b>11</b>	<b>Regression modeling</b>	<b>237</b>
11.1	Introduction . . . . .	237
11.2	Linear regression . . . . .	238
11.2.1	Scattered measurement values . . . . .	241
11.2.2	Discrete measurement values . . . . .	242
11.2.3	Uncertainty only exists in the response variable . . . . .	243
11.2.4	Modeling data that curves . . . . .	245
11.2.5	Visualizing the general trend . . . . .	248
11.2.6	Influential observations and Outliers . . . . .	249
11.2.7	Diagnosing problems in a regression model . . . . .	251
11.2.8	A model's goodness of fit . . . . .	252
11.2.9	Low signal-to-noise ratio . . . . .	253
11.2.10	Weighting data . . . . .	255

11.2.11 Sharp changes in a sequence of values . . . . .	255
11.3 Moving beyond the default Normal error . . . . .	256
11.3.1 Count data . . . . .	257
11.3.2 Continuous response variable having a lower bound . . . . .	259
11.3.3 Transforming the response variable . . . . .	259
11.3.4 Binary response variable . . . . .	261
11.3.5 Multinomial data . . . . .	262
11.3.6 Rates and proportions response variables . . . . .	262
11.3.7 Relational responses . . . . .	263
11.4 Multiple explanatory variables . . . . .	263
11.4.1 Interaction between variables . . . . .	267
11.4.2 Correlated explanatory variables . . . . .	268
11.4.3 Penalized regression . . . . .	272
11.5 Non-linear regression . . . . .	272
11.5.1 Power laws . . . . .	276
11.6 Mixed-effects models . . . . .	277
11.7 Generalised Additive Models . . . . .	280
11.8 Item response theory . . . . .	282
11.8.1 Agreement between raters . . . . .	282
11.9 Miscellaneous . . . . .	282
11.9.1 Advantages of using <code>lm</code> . . . . .	282
11.9.2 Network data . . . . .	283
11.9.3 Alternative residual metrics . . . . .	283
11.9.4 Quantized regression . . . . .	283
11.9.5 Prediction vs. interpretation . . . . .	283
11.9.6 Solving systems of equations . . . . .	283
11.9.7 Very large datasets . . . . .	284
11.9.8 Communicating model details . . . . .	284
11.10 Time series . . . . .	284
11.10.1 Cleaning time series data . . . . .	285
11.10.2 Modeling time series . . . . .	286
11.10.2.1 Building an ARMA model . . . . .	288
11.10.3 Non-constant variance . . . . .	291
11.10.4 Long-memory processes . . . . .	291
11.10.5 Smoothing and filtering . . . . .	291
11.10.6 Missing data . . . . .	292
11.10.7 Spectral analysis . . . . .	292
11.10.8 Relationships between time series . . . . .	292
11.10.9 Regression models . . . . .	293
11.10.10 Misc . . . . .	293
11.11 Survival analysis . . . . .	294
11.11.1 Kinds of censoring . . . . .	294
11.11.1.1 Input data format . . . . .	295

11.11.2 Survival curve . . . . .	295
11.11.3 Regression modeling . . . . .	297
11.11.3.1 Cox proportional-hazards model . . . . .	298
11.11.3.2 Time varying explanatory variables . . . . .	300
11.11.3.3 Parametric models . . . . .	303
11.11.4 Competing risks . . . . .	303
11.11.5 Multistate models . . . . .	304
11.12 Structural Equation Models . . . . .	304
11.13 Circular statistics . . . . .	304
11.13.1 Circular uniformity . . . . .	305
11.13.2 Fitting a regression model . . . . .	306
11.13.2.1 Linear response with a circular explanatory variable . . . . .	306
11.13.2.2 Circular response variable . . . . .	307
11.14 Compositions . . . . .	307
11.15 Extreme value statistics . . . . .	308
<b>12 Other techniques</b>	<b>309</b>
12.1 Machine learning . . . . .	309
12.1.1 Decision trees . . . . .	310
12.2 Clustering . . . . .	311
12.2.1 Principal component analysis . . . . .	312
12.2.2 Seriation . . . . .	312
12.3 Ordering of items . . . . .	314
12.4 Simulation . . . . .	315
12.5 Text analysis . . . . .	315
<b>13 Experiments</b>	<b>317</b>
13.1 Introduction . . . . .	317
13.1.1 Measurement uncertainty . . . . .	318
13.2 Design of experiments . . . . .	318
13.2.1 Subjects . . . . .	320
13.2.2 The task . . . . .	321
13.2.3 What is actually being measured? . . . . .	321
13.2.4 Stopping conditions . . . . .	322
13.2.5 Selecting experimental options . . . . .	323
13.3 Analysing the results . . . . .	324
13.3.1 Regression modeling . . . . .	325
13.3.2 Factorial designs . . . . .	326
13.3.3 Comparing sample means . . . . .	327
13.3.4 Comparing standard deviation . . . . .	332
13.3.5 Correlation . . . . .	332
13.3.6 Dichotomous variables . . . . .	333
13.3.7 Contingency tables . . . . .	334
13.3.8 ANOVA . . . . .	335

13.4 Benchmarking . . . . .	336
13.4.1 Following the herd . . . . .	337
13.4.2 Variability in today's computing systems . . . . .	337
13.4.2.1 Hardware variation . . . . .	338
13.4.2.2 Software variation . . . . .	342
13.4.2.3 End user systems . . . . .	344
13.4.2.4 The cloud . . . . .	345
13.5 User interface testing . . . . .	345
13.6 Surveys . . . . .	345
13.6.1 Checking survey reports . . . . .	347
<b>14 Overview of R</b>	<b>349</b>
14.1 Your first R program . . . . .	349
14.2 Language overview . . . . .	350
14.2.1 Differences between R and widely used languages . . . . .	350
14.2.2 Objects . . . . .	351
14.3 Operations on vectors . . . . .	352
14.3.1 Creating a vector/array/matrix . . . . .	352
14.3.2 Indexing . . . . .	352
14.3.3 Lists . . . . .	353
14.3.4 Data frames . . . . .	354
14.3.5 Symbolic forms . . . . .	355
14.3.6 Factors and levels . . . . .	355
14.4 Operators . . . . .	355
14.4.1 Testing for equality . . . . .	357
14.4.2 Assignment . . . . .	357
14.5 The R type (mode) system . . . . .	358
14.5.1 Converting the type (mode) of a value . . . . .	358
14.6 Statements . . . . .	358
14.7 Defining a function . . . . .	359
14.8 Commonly used functions . . . . .	359
14.9 Input/Output . . . . .	360
14.9.1 Graphical output . . . . .	360
14.10 Other uses for R . . . . .	361
14.11 Very large datasets . . . . .	361
14.12 Debugging R code . . . . .	361
<b>15 Data preparation</b>	<b>363</b>
15.1 Introduction . . . . .	363
15.1.1 Document cleaning operations . . . . .	364
15.2 Outliers . . . . .	365
15.3 Malformed file contents . . . . .	366
15.4 Missing data . . . . .	367
15.4.1 Handling missing values . . . . .	368

15.4.2 NA handling by library functions . . . . .	369
15.5 Restructuring data . . . . .	369
15.5.1 Reorganizing rows/columns . . . . .	370
15.6 Miscellaneous issues . . . . .	370
15.6.1 Application specific cleaning . . . . .	370
15.6.2 Different name, same meaning . . . . .	370
15.6.3 Multiple sources of signals . . . . .	371
15.6.4 Duplicate data . . . . .	371
15.6.5 Default values . . . . .	372
15.6.6 Resolution limit of measurements . . . . .	372
15.7 Detecting fabricated data . . . . .	372

# Read me 1st

This book aims to discuss all of what is currently known about software engineering, based on an analysis of all publicly available software engineering data.

This aim is not as ambitious as it sounds because there is not a great deal of data publicly available. Until recently researchers in software engineering concentrated on producing work that gave readers mathematical orgasms, rather than something that might be useful to industry based on experimental evidence.

As work on the book progressed, it became obvious that the best way to organise the material was as two parts, one covering software engineering and the second the statistics used in the analysis of software engineering data.

Life would have been easier if your author could have pointed readers at other books to learn about statistics. Unfortunately existing statistics books are not suitable, for reasons which include:

- they contain too much implementation detail. Developers are casual users of statistics and don't want to spend time learning lots of mathematics; they want to use the techniques, not implement them, and are only interested in the pre and post conditions,
- they target the largest market for introductory statistics books, the social sciences. The characteristics of the data encountered in social sciences are very different from the data encountered in software engineering, e.g., the Normal distribution is commonly encountered in social science data and is not that common in software engineering data, while the exponential distribution is common in software engineering and much less common in the social sciences,
- they continue to use statistical techniques that were designed to be practical for manual implementation (because electronic computers had not yet been invented). If fast computers are available, it is possible to use more powerful techniques which are impractical for manual implementation.

It is assumed that developer time is expensive and computer time is cheap. Where possible a single, general, statistical technique is described and a single way of coding something in R is used. This minimal, but general approach focuses on what developers need to know and the price paid is that the R code may be slower (in many cases the performance slowdown is unlikely to be noticeable).

The approach used is similar to that of a Doctor examining a patient for symptoms that can be matched against known underlying processes.

In some cases the questions I have asked about a particular set of measurements and the techniques used are very different from those made by the researchers who made the original measurements. Reasons for this include needing to illustrate a particular technique and making use of more powerful techniques to extract more information.

Some of the software engineering material, created on the first pass has the feel of a sequence of dots. A second pass is now underway to connect these dots into a coherent whole. The statistics for software engineering material has settled down and hopefully will be useful to people; there are some "...", "???" and places where the document production chain throws a wobbly.

The completed pdf will be made available for free online.

Formatting of large values on graph axis is courtesy of a time-machine from the 1970s; fixing this is on the TODO list.

If you have questions about empirical software engineering that this material does not answer, please let me know.

Even better, if you know of some interesting software engineering data, please tell me where I can download a copy.

All the code and data can be downloaded at: [github.com/Derek-Jones/ESEUR-code-data](https://github.com/Derek-Jones/ESEUR-code-data)

The names of data files usually share the same sequence of initial characters as the pdf file names of the corresponding research paper downloaded from the Internet.



# Chapter 1

## Introduction

Software systems and their host, the electronic computer,<sup>227</sup> began infiltrating the human ecosystem 70+ years ago.<sup>i</sup> Over this period the price of computer equipment continually declined, averaging 17.5% per year,<sup>1363</sup> an economic firestorm that devoured existing practices; software systems are apex predators. Figure 1.1 shows the dramatic fall in the cost of compute operations, however, without cheap mass storage computing would be a niche market; the continuing reduction in the cost of storage created increasing economic incentives for employing the cheap processing power; see Figure 1.2. The scale of improvements in hardware performance (and shortage of trained personnel) were appreciated almost from the beginning.<sup>557</sup>

A shift in perception, from computers as calculating machines,<sup>552</sup> to computing platforms, responding in real-time, to multiple independent users, created the opportunity to solve problems outside the realm machines dedicated to executing a single program, selected from a queue of waiting jobs. Figure 1.3 shows the initial growth of US based systems capable of sharing cpu time between multiple users.<sup>119</sup>

Since computers were first invented<sup>441,474</sup> people have learned how to do software development through personal experience, it is a craft activity. Most activities that build things start with people learning through their own and sometimes others' experience. In some cases what is being created is important enough that it becomes worth investing in research to discover effective engineering/scientific approaches to the various problems; the benefits of using such an engineering/scientific approach are greater control and predictability.

Why has software engineering not progressed from a craft to an engineering activity?

The engineering/scientific approach is based on theories that have been validated using empirical data. Until recently measurement data relating to the creation of software systems has either been unavailable or somewhat insubstantial. Over the last few years there has been an explosive growth in the collection and analysis of empirical data; it is now possible to write a book such as this one.

The lack of available data on commercial software development is a consequence of the longstanding relationship that has existed between customers and vendors. Change is driven by those with the power to make it happen and software systems has been a sellers market. The benefits of an engineering approach, are primarily reaped by customers, why should vendors invest in change if customers are willing to continue paying for systems developed using existing practices?

Those involved in building systems that use software want to control the process. Control requires understanding; understanding of the many processes involved in building software systems is the goal of software engineering research.

This is a data driven book that treats software engineering as an economic and cultural activity; it is intended to be useful to those involved in building software systems. A topic is only discussed if measurement data is publicly available to ground the discussion. Keeping to this requirement means that readers are likely to be dismayed at the scant coverage of many topics often covered at length in other books on software engineering.

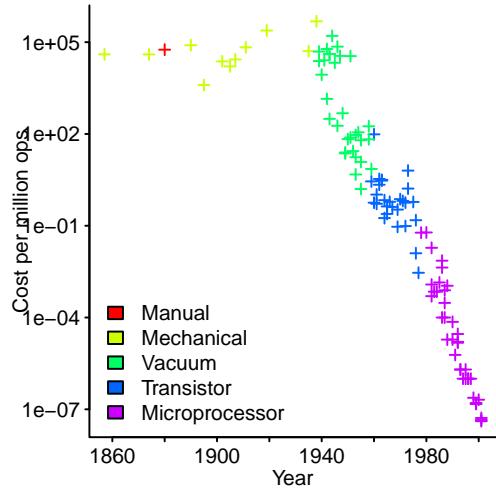


Figure 1.1: Total cost of one million computing operations over time. Data from Nordhaus.<sup>1018</sup> [code](#)

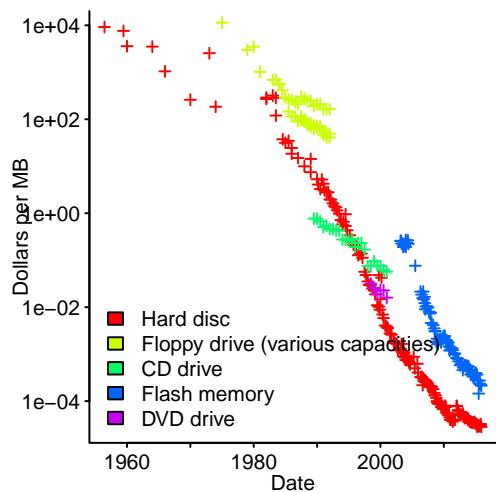


Figure 1.2: Storage cost, in US dollars per Mbyte, of mass market technologies over time. Data from McCallum,<sup>910</sup> floppy and CD-ROM data kindly provided by Davis.<sup>317</sup> [code](#)

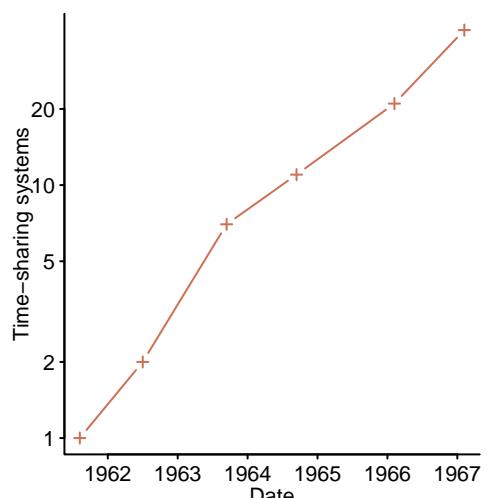


Figure 1.3: Initial growth of time-sharing systems available in the US. Data extracted from Gauthier.<sup>502</sup> [code](#)

<sup>i</sup> The verb "to program" was first used in its modern sense during 1946.<sup>551</sup>

Software is written within a particular development culture, by people having their own unique and changeable behavior patterns. Measurements of the products and processes in this environment are intrinsically noisy and are likely to include a variety of variables that are not measured. This does not mean that measurement and analysis is a futile activity, just that the uncertainty and variability is likely to be much larger than typically found in other engineering disciplines.

Measurement data is analysed using statistics, with statistical techniques being wielded as weaponised pattern recognition; those seeking a discussion of statistics whose purpose is to be a stimulant for mathematical orgasms, will not find satisfaction here.

Statistics does not assign a meaning to any of the patterns it uncovers; interpreting the patterns thrown up by statistical analysis, to give them meaning, is your job dear reader (based on your knowledge and experience of the problem domain).

The tool used for statistical analysis is the R system.<sup>1135</sup> R was chosen because of its ecosystem; there are many books, covering a wide range of subject areas, using R and active online forums discussing R usage (answers to problems can often be found by searching the Internet or if none are found a question can be posted with a reasonable likelihood of receiving an answer).

The data and R code used in this book are freely available for download from the book's website.<sup>705</sup>

Like programming, data analysis contains a craft component and the way to improve craft skills is to practice.

The continuing displacement of existing systems and practices has kept the focus of practice on development (of new systems and major enhancements to existing systems). A change of focus of practice to infrastructure<sup>1373</sup> has to await longer term stability.

?

## 1.1 Software ecosystems

The last 50 years or so has been a sellers market; the perceived benefits provided by software systems has been so significant that companies that did not use them risked being eclipsed by competitors. Whole industries were engulfed and companies became Red Queens, who had to keep running to maintain their position.

Provided software development projects looked like they would deliver something that was good enough, everybody knew that the customer would wait and pay; complaints could be ignored. Software vendors learned that the only way to survive in a rapidly evolving market was to get products to market quickly, before things moved on.

Experience from the introduction of earlier high-tech industries suggests that it takes many decades for major new technologies to settle down and reach market saturation.<sup>1076</sup> For instance, the transition from wood to steel for building battleships,<sup>1046</sup> started in 1858 and reached its zenith during the second world war; there is a long history of growth and decline various forms of infrastructure (see Figure 1.4).

Over the last 70 years a succession of companies have dominated the computing industry. The continual reduction in the cost of computing platforms created new markets and occasionally one of these grew to be the largest market, financially, for computing resources. A company dominates the computer industry when it dominates the market that dominates the industry. Companies that lose computing industry dominance often continue to grow, only declining when the market they have dominated, declines.

Figure 1.5 illustrates the impact of the growth of new markets on the market capitalization of three companies; IBM dominated when mainframes dominated the computer industry, the desktop market grew to dominate the computer industry and Microsoft dominated, smart phones removed the need for computers sitting on desks and these have grown to dominate the computer market with Apple being the largest company in this market (Google's Android investment is a defensive move to ensure they are not locked out of advertising on mobile, i.e., it is not as intended to be a direct source of revenue, making it extremely difficult to estimate its contribution to Google's market capitalization). It shows market capitalization

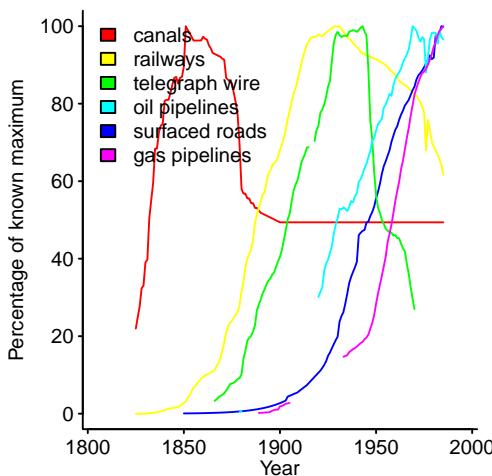


Figure 1.4: Growth of transport and product distribution infrastructure in the USA (underlying data is measured in miles). Data from Grübler et al.<sup>558</sup> [code](#)

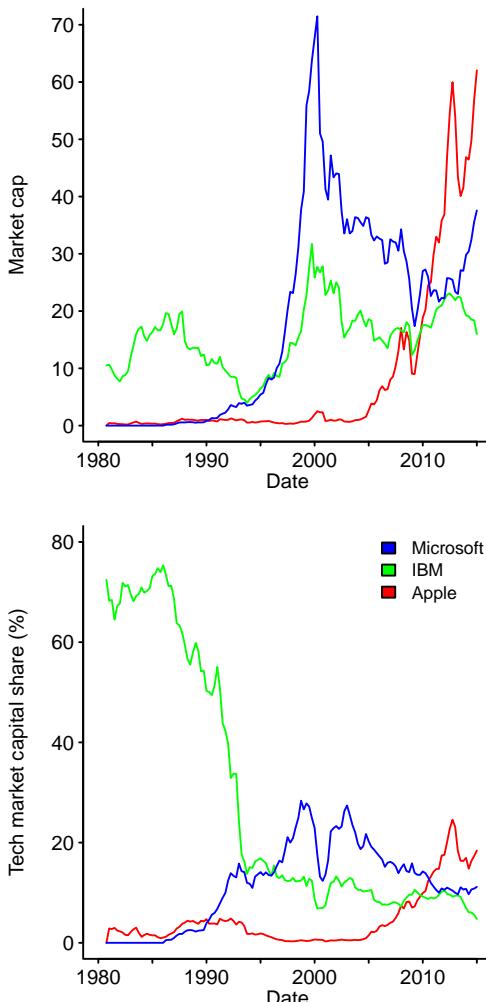


Figure 1.5: Market capitalization of IBM, Microsoft and Apple (upper), and expressed as a percentage of the top 100 listed US tech companies (lower). Data extracted from the Economist website.<sup>372</sup> [code](#)

(upper), and as a percentage of the top 100 listed US tech companies (lower), as of the first quarter of 2015.<sup>372</sup>

The three major eras, each with its own particular product and customer characteristics, have been (Figure 1.6 shows sales of major computing platforms):

- the IBM era, with mainframes as its main money spinner; high priced computers sold, or rented, to large organizations who either rented software from the hardware vendor or paid for software to be developed specifically for their own needs. In this ecosystem a single entity usually paid for software to be created, maintained and incurred the costs of any faulty operation of that software.

When the actual cost of software faults experienced by one organization is very high (with potential for even greater costs if things go wrong) and the same organization is paying all or most of the cost of creating the software, that organization can see a problem that is costing it money, that it thinks it should have control over. Very large organizations are in a position to influence research agendas to target the problems they want solved.

Large organizations tend to move slowly. The rate of change was slow enough for experience and knowledge of software engineering to be considered essential to do the job (this is not to say that anybody had to show that their skill was above some minimum level before they would be employed as a software developer).

- the Wintel era, with Microsoft Windows as the dominant software vendor and Intel the dominant hardware vendor; money to be made selling software packages to large numbers of customers. The direct cost of software maintenance is not visible to these customers, but they are paying for the costs of the consequences of faulty operation of that software.

Microsoft's mantra of a PC on every desk required that people write software to cover niche markets. The idea that anyone could create an application was promoted as a means of spreading Windows, by encouraging people with application domain knowledge to create software running under MS-DOS and later Windows.

Programming languages, libraries and user interface experienced high rates of change, which meant that developers found themselves on a learning treadmill. One lesson that many learned was that it was that the likelihood of change did not make it worthwhile investing in becoming an expert in a particular area.

- the Internet era, no single vendor dominates the Internet, but some large niches have dominant vendors, e.g., mobile phones,

It is difficult to judge whether the rate of change has been faster than in previous eras, or the volume of discussion about the changes has been higher because the changes have been visible to more people, or the lack of a dominant vendor to prevent change occurring too quickly.

The ongoing history of new software systems and computing platforms has created an environment where people are willing to invest their time and energy creating what they believe will be the next big thing. Those with the time and energy to do this tend to be young and inexperienced, outsiders in the sense that they don't have any implementation experience with existing systems. If any of these new systems take off, the developers involved will have made, or will make, many of the same mistakes made by the developers involved in earlier systems. The rate of decline of major software platforms is slow enough that employees with significant accumulated experience and expertise can continue to enjoy their seniority in well-paid jobs and have no incentive to jump ship to apply their expertise to an emerging system.

Mobile computing is only commercially feasible when the cost of computation, measured in Watts of electrical power, can be supplied by user-friendly portable batteries. Figure 1.7 shows the decline in electrical power consumed by a computation between 1946 and 2009; historically, it has been halving every 1.6 years.

Software systems have yet to reach a stable market equilibrium in many of the ecosystems they have colonised. Many software systems are still new enough that they are expected to adapt when the ecosystem in which they operate evolves. The operating characteristics of such systems have not yet been sufficiently absorbed into the fabric of life that they enjoy the power of making it easier to change the world to operate around them.

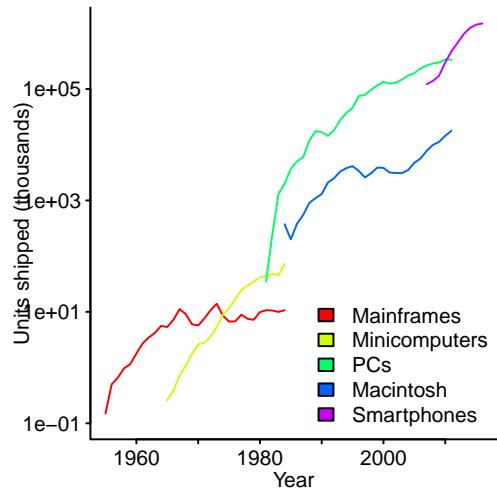


Figure 1.6: Total annual sales of computer species over the last 60 years. Data from Gordon<sup>524</sup> (mainframes and minicomputers), Reimer<sup>686</sup> (PCs) and Gartner<sup>480</sup> (smartphones). [code](#)

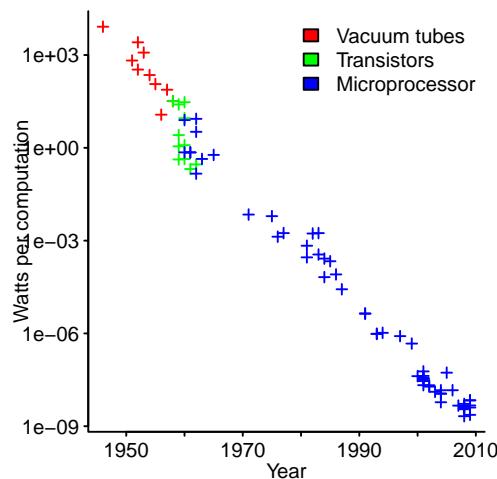


Figure 1.7: Power consumed, in Watts, executing an instruction on a computer available in a given year. Data from Koomey et al.<sup>781</sup> [code](#)

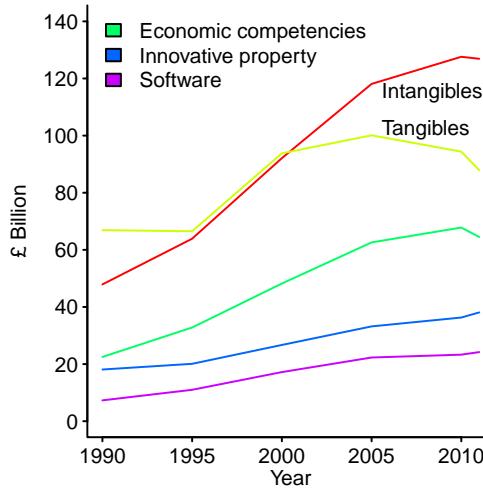


Figure 1.8: Total investment in tangible and intangible assets by UK companies, based on their audited accounts. Data from Goodridge et al.<sup>517</sup> [code](#)

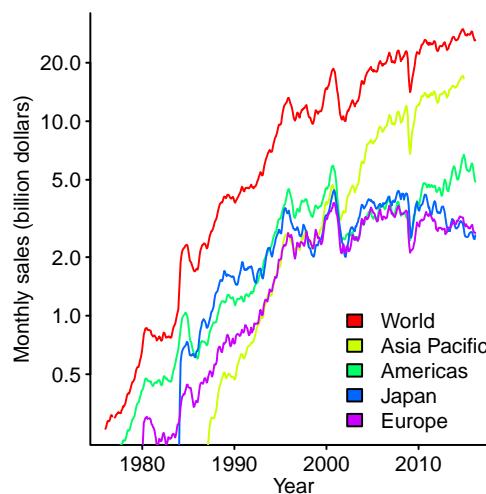


Figure 1.9: Billions of dollars of worldwide semiconductor sales per month. Data from World Semiconductor Trade Statistics.<sup>1464</sup> [code](#)

Economic activity is shifting towards being based around intangible goods;<sup>581</sup> cognitive capitalism is becoming mainstream.

A study by Goodridge, Haskel and Wallis<sup>517</sup> estimated the UK investment in intangible assets, as listed in the audited accounts that UK companies are required to file every year. Figure 1.8 shows the total tangible (e.g., buildings, machinery and computer hardware) and intangible assets between 1990 and 2012. Economic competencies are items such as training and branding, Innovative property includes scientific R&D, design and artistic originals (e.g., films, music and book); accounting issues associated with software development are discussed in Chapter 3.

A study by Wang<sup>1422</sup> found that while firms associated with the current IT fashion have a higher reputation and pay their executives more, they do not have higher performance. As companies selling hardware have discovered, designing products to become technologically obsolete (perceived or otherwise)<sup>1265</sup> or wear out and break after a few years,<sup>7</sup> creates a steady stream of sales. Given that software does not wear out, the desire to not to be seen as out of fashion provides a means for incentivizing users to update to the latest version.

Based on volume of semiconductor sales (see Figure 1.9), large new computer-based ecosystems are being created in Asia; the rest of the world appears to have completed their growth phase.

The economics of chip fabrication,<sup>792</sup> from which Moore's law derived,<sup>ii</sup> was what made the firestorm possible; the ability to create more products (by shrinking the size of the components; see Figure 1.10) for roughly the same production cost (of going through the chip fabrication process),<sup>633</sup> faster processors and increased storage capacity on similar-sized devices were fortuitous side effects. The upward ramp of the logistic equation has now levelled off<sup>442</sup> and by the time this book appears Moore's law will be something that people used to write about.

Software systems are part of a world-wide economic system that has been rapidly evolving for several hundred years; the factors driving economic growth are slowly starting to be understood.<sup>695</sup> Analysis of changes in World GDP have found cycles, or waves, of economic activity; Kondratieff waves are the longest, with a period of around 50 years (data from more centuries may contain a longer cycle), a variety of shorter cycles are also seen, such as the Kuznets swing of around 20 years. Five Kondratieff waves have been identified with major world economic cycles, e.g., from the industrial revolution to and information technology. Figure 1.11 shows a spectral analysis of estimated World GDP between 1870 and 2008; adjusting for the two World-wars produces a smoother result.

Computers were the enablers for the latest wave, from the electronic century.

### 1.1.1 What are the software engineering activities?

Software engineering is the collection of activities performed by people involved in producing and maintaining software executed by computer to solve a problem.

These activities will have some path dependency, that is, once the know-how and infrastructure for performing some activity becomes established the existing practice is likely to continue.

Perhaps the most entrenched path dependency in software development is the use of two-valued logic, i.e., binary. The most efficient radix, in terms of representation space (number of digits times number of possible values of each digit), is,  $2.718\dots$ ,<sup>592</sup> the closest integral value is 3. The use of binary, rather than ternary, was driven by the characteristics of the available electronic switching devices.<sup>iii</sup> Given the vast quantity of software making an implicit, and in some cases explicit, assumption that binary representation is used, a future switching technology that would support the use of a ternary representation might not be adopted or be limited to resource constrained environments.<sup>1429</sup>

<sup>ii</sup> Moore's original paper,<sup>969</sup> published in 1965, extrapolated four data-points to 1975 and questioned whether it would be technically possible to continue the trend

<sup>iii</sup> In a transistor switch, Off is represented by very low-voltage/high-current and On represented by saturated high-voltage/very low-current. Transistors in these two states consume very little power (power equals voltage times current). A third state would have to be represented at a voltage/current point that would consume significantly more power. Power consumption, or rather the heat generated by it, is a significant limiting factor in processors built using transistors.

Traditionally activities have included: obtaining requirements, creating specifications, design at all levels, writing and maintaining code, writing manuals fixing problems and providing user support. Employees within large organizations specialise within a particular area and these activities are the areas where software developers specialise.

In small companies there is greater opportunity, and sometimes a need, for employees to become involved in tasks that would not be considered part of the job of a software developer in larger companies. For instance, being involved in any or all of a company's activities from the initial sales inquiry through to customer support of the delivered system; the financial aspect of running a business is likely to be much more visible in a small company.

Software is created and used within a variety of ecosystems and software engineering activities can only be understood in the context of the ecosystem in which it operates.

While this is an overly broad definition, let's be ambitious and run with it, allowing the constraints of data availability and completing a book to provide the filtering.

...<sup>117</sup>

The debate over the identity of computing as an academic discipline is ongoing.<sup>1341</sup>

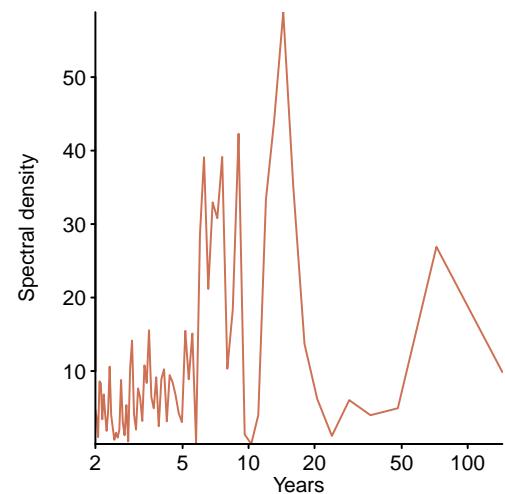


Figure 1.11: Spectral analysis of World GDP between 1870-2008; peaks around 17 and 70 years. Data from Maddison.<sup>885</sup> [code](#)

## 1.2 Brief history of software engineering research

The fact that software often contained many faults and took much longer than expected to produce was a surprise to those working at the start of electronic computing, after World War II. Established practices for measuring and documenting the performance of electronics were in place and ready to be used for computer hardware.<sup>774, 1087</sup> It was not until the end of the 1960s that a summary of the major issues appeared in print.<sup>995</sup>

Until the early 1980s most software systems were developed for large organizations and over 50% of US government research funding for mathematics and computer science came from the Department of Defense,<sup>440</sup> an organization that built large systems over time-frames of many years.

Large organizations, such as the DOD, spend so much on software that it is worth their while investing in research aimed at reducing costs and learning how to better control the development process. During the 1970s project data, funding and management by the Rome Air Development Center,<sup>iv</sup> RADC, came together to produce the first collection of wide-ranging empirically based reports analysing the factors involved in the development of large software systems.<sup>1349, 1379</sup>

For whatever reason the data available at RADC was not widely distributed or even known about; the only people making use of this data in the 1980s and 1990s appear to be Air Force officers writing Master's thesis.<sup>1042, 1249</sup>

The legacy of this first 30 years was a research agenda oriented towards building, in their day, very large systems.

Most published software engineering research since around 1980 has not made use of empirical evidence; unless empirical research does not include theoretical contribution it may fail to be accepted for publication.<sup>10</sup> A review<sup>869</sup> in the early 1990s, of published papers relating to software engineering, found an almost total lack of empirical analysis of its engineering characteristics; a systematic review of 5,453 papers published between 1993 and 2002<sup>579</sup> found 2% reporting experiments. When experiments had been performed, they suffered from small sample sizes<sup>729</sup> (a review<sup>1418</sup> using papers from 2005 found that little had changed), had statistical power falling well below norms used in other disciplines<sup>369</sup> or simply failed to report an effect size (for the 92 controlled experiments published between 1993 and 2002 only 29% reported an effect size<sup>729</sup>).

Why have academics working in an engineering discipline not based their research on empirical data? Perhaps the difficulty of obtaining realistic data<sup>1118</sup> was an important factor (commercial companies are loath to have outsiders measuring what they are doing and many do not measure themselves, so even confidential is hard to find). The lack of data discouraged anybody wanting to do intellectually sound research, with those investigating the field learning how difficult it is to make worthwhile progress without reliable data (a few intrepid souls

<sup>iv</sup> The main US Air Force research lab. There is probably more software engineering data to be found in US Air Force officers' Master's thesis, than all academic software engineering papers published before the early 2000s.

did run experiments using professional developers<sup>100</sup>). Researchers with a talent for software engineering either moved on to other research areas or to work in industry, leaving the field to those with talents in the less employable areas of mathematical theory, literary criticism (of source code) or folklore.<sup>190</sup>

A lack of data has not prevented researchers expounding plausible sounding theories that, in some cases, have become widely regarded as true. For instance, it was once claimed, without any empirical evidence, that the use of source code clones (i.e., copying code from another part of the project) is bad practice (e.g., clones are likely to be a source of faults, perhaps because only one of the copies was updated).<sup>456</sup> In practice research has shown that the opposite is true,<sup>1142,1355</sup> clones are less likely to contain faults than *uncloned* source.

Given the sparsity of available empirical data relating to industrial software systems, it is no surprise that academics researchers have failed to create any reliable predictive or descriptive software engineering theories having a connection to reality.<sup>v</sup>

Many of the theories relating to software engineering processes commonly encountered in academic publications are based on ideas created many years ago by somebody who was able to gain access to a relevant (often tiny) dataset. For instance, Perry<sup>1077</sup> divided software interface faults into 15 categories using a data set of just 85 modification requests to draw conclusions; this work is still being cited in papers 25 years later. These fossil theories have continued to exist because of the sparsity of data needed to refute or improve on them.

Human psychology and sociology continue to be completely ignored as major topics of software research, a fact pointed out over 35 years ago.<sup>1238</sup>

Quantity of published papers should not be confused with progress towards an effective model of behavior. For instance there has been a great deal published on software process improvement, SPI, (635 papers were found by a recent study<sup>790</sup> of research over the last 25 years), with experience reports and proposed solutions making up two-thirds of the publications; however, the proposed solutions were barely evaluated, there were no studies evaluating advantages and disadvantages of proposals and the few testable theories are waiting to be tested.

Over the last 10 years or so there has been an explosion of empirical research, driven by the availability of large amounts of data extracted from open source software. This significant change in the availability of data and the resulting boom in the fortunes of empirical research does not mean that existing theories and ideas will change overnight. Simply ignoring all research published before 2005 (roughly when the data deluge started) does not help, earlier research has seeded old wives tales that have become embedded in the folklore of software engineering creating a legacy that is likely to be with us for sometime to come.

It is inevitable that some early papers will make claims about software engineering that we now believe to be true. It is possible to search though the contents of old papers looking for wording that can be interpreted as a connection with a recent discovery in the same way that it is possible to search through the prophecies of Nostradamus to find one that can be interpreted as predicting the same discovery.

This book takes the approach that empirically verified theories in software engineering don't yet exist, the subject is a blank slate. Knowing that certain patterns of behavior regularly occur is an empirical observation, a theory would make verifiable predictions that include the observed patterns. In some cases existing old wives tales are discussed if it is felt that their use in an engineering environment would be seriously counter-productive. For instance, while various software metrics (e.g., Halstead's metric) are widely known, your authors' experience is that practicing developers do not invest effort in using them; they are famous for being famous and so there is little to be gained in spending much effort debunking them.

### 1.2.1 The academic ecosystem

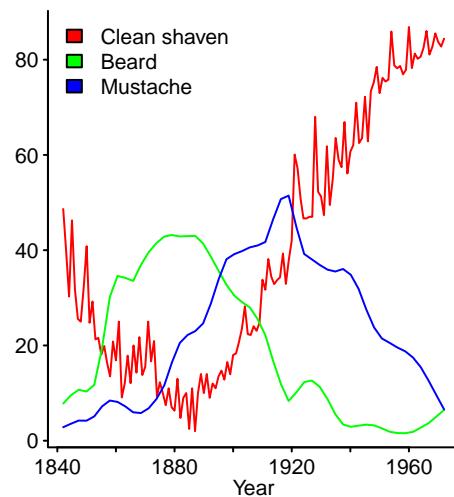
Interactions between people in industry and academia has often suffered from a misunderstanding of each other's motivations and career pressures.

Few people in industry have much interaction with the academic ecosystem that claims to research their field. The quaint image of researchers toiling away for years before publishing a carefully crafted manuscript is long gone.<sup>376</sup> Although academics continue to work in a

---

<sup>v</sup> The data that was available was generally very small, with little supporting context. See (am I really going to create this collection???) for a collection of historical data sets.

Figure 1.12: Changing habits in men's facial hair. Data from Robinson.<sup>1166</sup> code



feudal based system of patronage and reputation, they are incentivised by the motto “publish or perish”,<sup>1063</sup> with science perhaps advancing one funeral at a time.<sup>68</sup> Hopefully the change to empirically based software engineering research will progress faster than fashions in men’s facial hair (most academics researching software engineering are men); see Figure 1.12.

Academic research projects share many of the characteristics of commercial startups. They involve a few people attempting to solve a sometimes fuzzily defined problem, trying to make an improvement in one area of an existing *product* and they often fail, with the few that succeed producing spectacular returns. Researchers are serial entrepreneurs in that they tend to only work on funded projects, moving onto other projects when funding runs out (and often having little interest in previous projects). Like commercial product development, the choice of research topics is fashion driven; see Figure 1.13.

The visible output from academic research are papers published in journals and conference proceedings. It is important to remember that: “ . . . an article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”<sup>195</sup>

Many journals and conferences uses a process known as *peer review* to decide whether a submitted paper should be accepted. The peer review process involves sending submitted papers to a few referees, ideally chosen for their expertise of the topic covered, who make suggestions on how to improve the paper and provide a yes/no acceptance decision (the identity of the paper’s author(s) and the referees are often anonymous). The peer review process has generally worked well in the past, but is looking to be in need of a major overhaul<sup>1116</sup> to improve turn-around time and handle the work load generated by a proliferation of journals and conferences; some journals and conferences have become known for the low quality of their peer reviews. The status attached to the peer review process has resulted in it being adopted by journals covering non-traditional fields, e.g., psychic research.

In the past most researchers have made little, if any, effort to archive the data they gather for future use. One study<sup>1411</sup> requested the datasets relating to 516 biology related studies, with ages between 2 and 22 years; they found that the probability of the dataset being available fell by 17% per year. Your author’s experience of requesting data from researchers is that it often fails to survive beyond the lifetime of the computer on which it was originally held.

Researchers may have reasons, other than carelessness, for not making their data generally available. For instance, a study<sup>1447</sup> of 49 papers in major psychology journals found that the weaker the evidence for the researcher’s hypothesis the less likely they were to be willing to share their data.

The importance of making code and data available is becoming widely acknowledged, with a growing number of individual researchers making code/data available for download and journals having explicit policies about authors making this information available.<sup>1317</sup> In the UK researchers who receive any funding from a government research body are now required to archive their data and make it generally available.<sup>1155</sup>

Discovering an interesting pattern in experimental data is the start, not the end of experimentation involving that pattern (the likelihood of obtaining a positive result is as much to do with the subject studied as the question being asked<sup>408</sup>). The more often a behavior is experimentally observed the greater the belief that the effect seem actually exists. Replication is the process of duplicating the experiment(s) described in a report or paper to confirm the findings previously obtained. For replication to be practical researchers need to be willing to share their code and data, something that a growing numbers of software engineering researchers are starting to do; those attempting replication are meeting with mixed success.<sup>276</sup>

Replication is not a high status activity, with high impact journals interested in publishing research which reports new experimental findings and not wanting to dilute their high impact status by publishing replications (which, when they are performed and fail to replicate previous results considered to be important discoveries, can often only find a home for publication in a low impact journal). A study<sup>1034</sup> replicating 100 psychology experiments found that while 97% of the original papers reported p-values less than 0.05, only 36% of the replicated experiments obtained p-values this low; a replication study<sup>232</sup> of 67 economics papers was able to replicate 22 (33%) without assistance from the authors, 29 with assistance.

Replication is necessary, i.e., the results claimed by one researcher must be reproduced by another if they are to be accepted. Without replication researchers are amassing a graveyard of undead theories.<sup>426</sup>

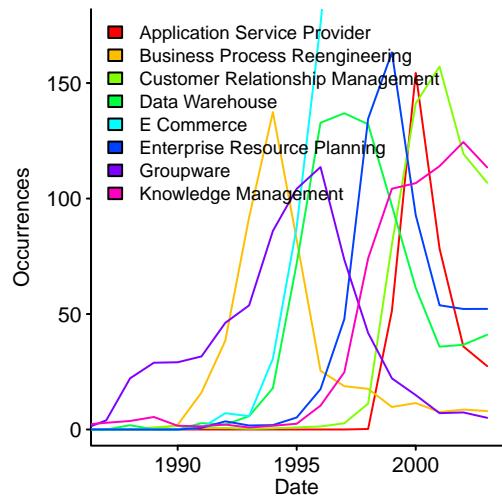


Figure 1.13: Number of papers, in each year between 1987 and 2003, associated with a particular IT topic. The E-commerce paper count peaks at 1,775 in 2000 and in 2003 is still off the scale compared to other topics. Data kindly provided by Wang.<sup>1422</sup> code

These days academic performance is often measured by number of papers published and the impact factor of the journal in which they were published<sup>794</sup> (scientific journals publish a percentage of the papers submitted to them, with prestigious high impact journals publishing a lower percentage than those with lower impact factors; journals and clubs share a common economic model<sup>110</sup>). Organizations that award grants to researchers often consider the number of published papers and impact factor of the publication journal when deciding whether to award a grant; the effect is to generate an evolutionary pressure that selects for bad science.<sup>1274</sup>

One consequence of the important role of published paper count in an academic's career is an increase in scientific fraud,<sup>407</sup> most of which goes undetected;<sup>273</sup> one study<sup>409</sup> found that most retracted papers (67.4%) were attributable to misconduct, around a 10-fold increase since 1975. More highly ranked journals have also been found to contain a higher percentage of retracted papers,<sup>177</sup> it is not known whether this represents an increased in flawed articles or an increase in detection.<sup>1304</sup> Only a handful of software engineering papers have been retracted, perhaps because a lack of data makes it very difficult to verify any of the claims made by the authors. The website [retractionwatch.com](http://retractionwatch.com) reports on possible and actual paper retractions.

**Commercial research labs** Many large companies have research groups. While the researchers working in these groups often attend the same conferences as academics and publish papers in the same journals, their performance is often measured by the number of patents granted. The number of software patents continues to grow, with 109,281 granted in 2014<sup>1015</sup> (36% of all utility patents).

### 1.2.2 Paper citation practices

This book attempts to provide citations to any claims made, so that readers can perform background checks. To be cited papers have to be freely available for public download, unless published before 2000 (or so), and when data is analysed it has to be free for public distribution.<sup>vi</sup>

When academics claim that their papers can be freely downloaded what they may mean is that the university that employs them has paid for a site-wide subscription that enables university employees to download copies of papers published by various commercial publishers. Tax payers pay for the research and university subscriptions and may not have free access to it. Things are slowly changing. In the UK researchers in receipt of government funding are now incentivized to publish in journals that will make papers produced by this research freely available after a period of 6-12 months from publication date. Your author's attitude is that academics are funded by tax payers and if they are unwilling to provide a freely downloadable copy on their web page, they should not receive any credit for the work.<sup>vii</sup>

Software developers are accustomed to treating documents that are more than a few years old as being out-of-date; a consequence of the fast changing environment in which they operate. Some fields, such as cognitive psychology, are more established and people do not feel the need to keep repeating work that was first performed decades ago (although it might be replicated as part of the process of testing new hypothesis). In more established fields it is counter-productive to treat date of publication as a worthiness indicator.

## 1.3 Lessons learned from the analysis in this book

The gold standard of the scientific method is the controlled randomised experiment, followed by replication of the results by others (failure to replicate the finding of the original study is common<sup>665</sup>). All the factors that could influence the outcome of an experiment are either controlled or randomly divided up such that any unknown effects add noise to the signal rather than ghost patterns.

A handful of the 310? datasets analysed in this book came from controlled experiments, some of which used randomization, and a handful of this handful have been replicated.

---

<sup>vi</sup> The usual academic procedure is to cite the first paper that proposes a new theory, describes a pattern of behavior, etc.

<sup>vii</sup> In fact they should not have been funded in the first place; if an academic refuses to make a copy of their papers freely available to you please report their behavior to your elected representative.

Open source is readily available in quantity and a lot of empirical research now makes use of this public resource. To what extent are results obtained from measurements on Open Source projects applicable to non-Open Source software development, i.e., how closely do the characteristics of open source match the characteristics of closed source? In some cases there may be little reason to expect any differences, while in others large differences may exist (e.g., the speed with which developers respond to security advisories<sup>791</sup> is likely to be faster in a commercial environment). The only definitive answer comes from comparing data from both environments.

Other possible differences might arise from Open Source software developers having a larger say in what features are implemented, rather than the manager or individual clients; successful Open Source projects probably have many more users than most commercial systems (which may only be used internally or have a few large customers)...

The patterns of behavior seen in the analysis discussed in this book should be treated as suggestions for what might be. Perhaps their most practical application is in helping to dismiss behaviors that might otherwise be thought possible.

and...

The measurement process invariably introduces some amount of error. While it may be possible to measure source code to a high degree of accuracy, other the values of software related quantities are likely to be a lot less certain. Table 1.1 lists estimates of the value of U.S. Computer manufacturer shipments, made by different data gathering organizations.

Authority	Worldwide			Domestic		
	1960	1965	1970	1960	1965	1970
EIA	0.630	2.830	5.162		2.575	3.958
EDP/IR	0.72	2.40	7.29	0.59	1.77	4.37
ADL			7.27	0.53	2.10	4.94

Table 1.1: Estimated value of shipments by U.S. computer manufacturers, in \$billion, made by various data gathering organizations. Data from Phister<sup>1087</sup> table 0.2.

## 1.4 Overview of contents

This book has two parts:

- a discussion of the major components of software engineering, driven by an analysis of publicly available data. The intent is to cover the publicly available evidence, enumerating patterns of behavior that have been found in software engineering activities. Many topics that are usually covered in software engineering textbooks are not discussed because public data relating to them could not be found,
- an illustration of statistical techniques applicable to the kinds of measurement data and problems likely to be encountered by professional software developers; this provides the foundation for the statistical analysis in the first part.

The intent is to provide pointers to statistical and experimental techniques that might be used to help analyse a sample of measurements or provide guidance for solving a data driven problem in software engineering.

While readers will learn something about statistics the material is not intended to provide an introduction to statistics as such, but rather an introduction to the use of statistical concepts and methods, along with the assumptions that underlie them. A suggested reading list is provided...

**Human cognitive characteristics** The operating characteristics of the human brain is an essential part of any study of software engineering. Characteristics such as cognitive capacity, memory storage/recall performance and learning, personality (e.g., propensity to take risks) and processing of visual information are discussed.

Software developers come preloaded with overlearned behaviors, derived from the native culture of their formative years and at least one human language which they have spent many hours reading

**Cognitive capitalism** Current economic theory and practice is predominantly based around the use of labor as the means of production; the economics of products produced by intellectual effort has been growing in importance for some time (the chapter title provides a direction for how to think about the economic material). The approach to software economics is from the perspective of the software engineer or software company rather than the perspective of the customer or user of software (which differs from much existing work which adopts the customer or user perspective).

Introductory economic issues are discussed.

**Ecosystems** Software is created in a development ecosystem that has been built and maintained by many people, and is used in a customer ecosystem that generally involves many people. Once written software is unchanging, if the world in which the software is used changes there will be pressure to change the software; software systems only change because the world in which they are used changes and creates an incentive to make an investment in changing the software.

Various components of developer (e.g., careers) and development (e.g., APIs) ecosystems are discussed, along with non-software issues having some form of direct impact on developers.

**Projects** The issues and stages of building a software system, from bidding to implement a system to delivery and ongoing maintenance of the system.

**Reliability** How can the reliability of a software system be estimated and what is the contribution made by various components of a system? Why and where are faults experienced and when are they worth fixing? Economics and marketing drive the effort invested in searching for and removing mistakes present in the code, as well as deciding which mistakes are an acceptable risk in a shipped product.

**Source code** Patterns of human and programming language usage in source code. What information do these patterns provide about the habits and characteristics of the people who produced the code?

**Stories told by data** There is no point analysing data unless the results can be effectively communicated to the intended audience. Possible techniques for communication a story found in data are covered, along with issues to look out for in stories told by others.

**Probability** An overview involving basic principles in probability, along with commonly used constructs, e.g., probability distributions, means and variance, permutation and randomizations tests.

**Statistics for software engineering** No statistical knowledge is assumed on the part of the reader, however, it is assumed that readers have basic algebra skills and can interpret graphs.

The approach used is similar to that of a Doctor examining a patient for symptoms that provide pointers to underlying processes. Developers are casual users of statistics and don't want to spend time learning lots of mathematics; they want to make use of techniques, not implement them.

In many practical situations the most useful expertise to have is knowledge of the application domain that generated the data and how the results might be applied.

It is better to calculate an approximate answer to the correct problem than an exact answer to the wrong problem.

Because evidence-based software engineering is only recently started to be applied, there is uncertainty about which statistical techniques are most likely to be generally applicable. Therefore, an all encompassing approach is taken and a broad spectrum of topics are covered, including:

- statistical concepts: sampling, describing data, p-value, confidence intervals, effect size, statistical power, bootstrapping, model building, comparing two or more groups,
- regression modeling: fitting data to an equation can provide information about which variables have the most power to explain measurable behavior. Software engineering measurements come in a wide variety of forms, while ordinary least squares might be widely used in the social sciences, it is not always suitable for modeling the datasets encountered in software engineering; more powerful techniques such as generalized least squares, nonlinear models, mixed models, additive models, structural equation models and others sometimes have to be used,

- time series: analysis of data where measurements at time  $t$  are correlated with measurements made at past times, e.g.,  $t - 1$ , is the domain of time series analysis (regression modeling assumes that successive measurements are independent of each other),
- survival analysis: measurements of time to an event occurring (e.g., death) are the domain of survival analysis,
- machine learning: various techniques for finding patterns in data when the person doing the analysis has no knowledge of the data or the application domain that produced it. Sometimes we are clueless button pushers and machine learning can be a useful guide.

### Experiments

General issues involving the design of experiments are discussed.

Probably the most common experiment performed by developers is hardware and software benchmarking. The many difficulties and complications involved in performing reliable benchmarks are illustrated.

Another form of experiment is product testing, e.g., website design. In a small company a software developer may have a lot of influence on the company website and knowing how to test the effectiveness of different designs is important.

**Surveys:** Analysis of data obtained by asking people questions.

**Data cleaning** Garbage in garbage out. Data cleaning is the penultimate chapter (and goes unmentioned in some books), and is often the most time-consuming part of data analysis and a very necessary activity for obtaining reliable results.

Common data cleaning tasks, along with possible techniques for detecting potential problems and solving them using R are discussed.

**Overview of R** An overview of R aimed at developers who are fluent in at least one other computer language. The discussion concentrates on those language features likely to be commonly used, but behave surprisingly differently from languages the reader is likely to be familiar with.

**Obtaining and installing R:** If you cannot figure out how to obtain and install R, this book is not for you.

RStudio is a widely used R IDE that is also sometimes used by your author.

## 1.4.1 Why use R?

The main reasons for selecting R as the language+support library in which to write the statistical analysis programs in this book were:

- it is possible to quickly write a short program that solves the kind of problems that often occur when analysing software engineering data. The process often follows the sequence: read data from one of a wide variety of sources, operate on it using functions selected from a huge library of existing packages and finally graphically displaying the results or printing values,
- lots of people are using it: there is a very active ecosystem with many R books, active discussion forums where examples can be found, answers to old questions read and new questions posted,
- accuracy and reliability: a comparison of the reliability of 10 statistical software packages<sup>1028</sup> found that GAUSS, LIMDEP, Mathematica, MATLAB, R, SAS, and Stata provided consistent reliable estimation results, and a comparison of the statistical functions in Octave, Python, and R<sup>24</sup> found that R yielded the best results. A study<sup>25</sup> of the precision of five spreadsheets (Calc, Excel, Gnumeric, NeoOffice and Oleo) running under Windows Vista, Ubuntu and Mac OS found that no one spreadsheet provided consistently good results (it was recommended that none of these spreadsheets be used for nonlinear regression and/or Monte Carlo simulation); another study<sup>917</sup> found significant errors in the accuracy of the statistical procedures in various versions of Microsoft Excel.
- an extensive library of add-on packages (over 7,000 at the time of writing): CRAN, the Comprehensive R Archive Network is the official package library; some packages are only available on R-Forge and Github.

A freely available open source implementation is always nice to have.

## 1.5 Terminology, concepts and notation

Much of the basic terminology used in probability and statistics in common use today derives from gambling and experimental research in medicine and agriculture, because these were the domains where researchers working in the early days of statistics were employed.

- *group*: each sample is sometimes referred to as a group,
- *treatment*: The operation or process performed is often referred to as a treatment.
- *response variable* also known as a *dependent variable*, responds/depends to/on changes of values of explanatory variables; their behavior depends on the variables that are independent (at least of them),
- *explanatory variables* (also known as *independent, stimulus* or *predictor variables*), are used to explain, predict or stimulate the value of response variables; they are independent of the response variable,
- Data is *truncated* when values below or above some threshold are unobserved (or removed from the dataset).
- Data is *censored* when values below or above some threshold are set equal to the threshold.

*between subjects*: when samples are obtained from different groups of subjects, often with the different groups performing a task under different experimental conditions,

*within subjects*: Comparing two or more samples obtaining using the same group of subjects, often with the different subjects performing a task under two or more different experimental conditions,

Some commonly encountered symbols and notation:

- $n!$  ( $n$  factorial), denotes the expression  $n(n - 1)(n - 2) \cdots 1$ , available in the `factorial` function,
- $\binom{n}{r} = \frac{n!}{r!(n-r)!}$ , a commonly occurring quantity in the analysis of probability problems; calculated by the `choose` function,
- a hat above a variable,  $\hat{y}$ , denotes an estimate, in this case an estimate of  $y$ 's value,
- $\mu$  (mu), commonly denotes the mean value, available in the `mean` function,
- $\sigma$  (sigma), commonly denotes the standard deviation, available in the `sd` function. The terms 1-sigma, 2-sigma, etc are sometimes used to refer to the probability of an event occurring. Figure 1.14 shows the sigma multiplier for various probabilities,
- $n \rightarrow \infty$  as  $n$  goes to infinity, i.e., becomes very very large,
- $n \rightarrow 0$ , as  $n$  goes to zero, i.e., becomes very very small,
- $P(x)$ , the probability of  $x$  occurring and sometimes used to denote the Poisson distribution with parameter  $x$  (however, this case is usually written using  $\lambda$ , e.g.,  $P(\lambda)$ ),
- $P(a < X)$ , the probability that  $a < X$ . The functions `pnorm`, `pbinom` and `ppois` can be used to obtain the probability of encountering a value less than or equal to  $x$  for the respective distribution (e.g., Normal, Binomial and Poisson, as suggested by the naming convention),
- $P(|a - X|)$ , the probability of the absolute value of the difference between  $a$  and  $X$ ,

- $\prod_{i=1}^6 a_i$ , the product:  $a_1 \times a_2 \times \cdots \times a_6$ ,
- $\sum_{i=1}^6 P(a_i)$ , the sum:  $P(a_1) + P(a_2) + \cdots + P(a_6)$ ,

The sum the probabilities of all the mutually exclusive things that could happen, when an action occurs, is always one. For instance, when a die is rolled the six probabilities of a particular number occurring sum to one, irrespective of whether the die is fair or has been tampered with in some way.

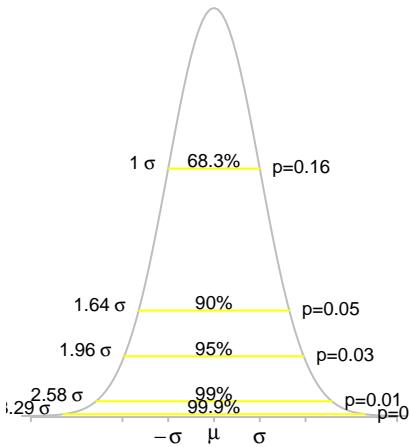


Figure 1.14: Normal distribution with total percentage of values enclosed within a given number of standard deviations. [code](#)

- $P(D|S)$ , the probability of  $D$  occurring given that  $S$  is true; known as the *conditional probability*. For instance,  $S$  might be the event that two dice have been rolled and their face-up numbers sum to five and  $D$  the event that the value of the first die is four.

The value can be calculated using the following:

$$P(D|S) = \frac{P(DS)}{P(S)}$$

where  $P(DS)$  is the probability of both  $D$  and  $S$  occurring together.

If  $D$  and  $S$  are independent of each other, then we have:

$$P(DS) = P(D)P(S)$$

and the above equation simplifies to:  $P(D|S) = P(DS)$

A lot of the existing theory in statistics assumes that variables are independent. Characteristics unique to each dice means using a different dice for each roll will produce values having slightly more independence than using the same dice twice.

**Convex/concave functions:** a function  $f(x)$  is *convex*, between  $a$  and  $b$ , if every chord of the function is above the function. If the chord is always below the function, it is concave; see Figure 1.15. The word convex tends to be spoken with a smiley face, while concave induces more of a frown.

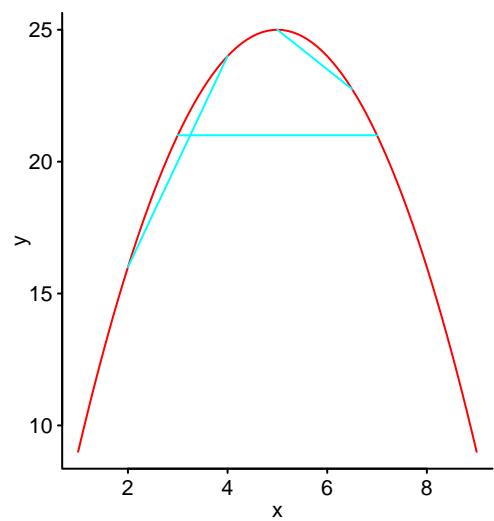
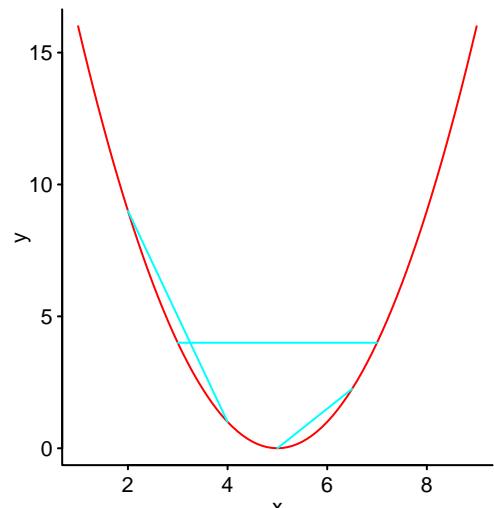


Figure 1.15: Example convex, upper, and concave, lower, functions; lines are three chords of the function. [code](#)



# Chapter 2

## Human cognitive characteristics

### 2.1 Introduction

Software systems are built and maintained by the creative output of the human brain, which supplies the cognitive effort that directs and performs the activities required; the operating characteristics of this machine are an essential component of any study of software engineering.

Modern humans evolved from earlier humanoids who in turn evolved from earlier members of the ape family who in turn evolved from etc., etc. The collection of cognitive characteristics present in homo sapien brains is the end-result of a particular sequence of survival requirements that occurred over millions of years of evolutionary history; with the last common ancestor of the great apes and the line leading to modern humans living 5 to 7 million years ago,<sup>452</sup> the last few hundred thousand years spent as hunter-gatherers roaming the African savannah, followed by 10,000 years or so having a lifestyle that involved farming crops and raising domesticated animals.

Our skull houses a computing system that evolved to provide responses to problems that occurred in the stone age ecosystem. However, neural circuits in the brain established for one purpose may be redeployed,<sup>40</sup> during normal development, for different uses, often without losing their original functions, e.g., in many people learning to read and write involves repurposing the neurons in the ventral visual occipito-temporal cortex (an area of the brain involved in face processing and mirror-invariant visual recognition).<sup>333</sup> Reuse of neural circuitry is a central organizational principle of the brain.

The collection of cognitive characteristics supported by an animal's brain only makes sense in the context of the problems it had to solve in the environment in which it evolved. Cognition and the environment are like the two blades of a pair of scissors (Figure 2.1), both blades have to mesh together to achieve the desired result.

- The structure of the natural environment places constraints on optimal performance (an approach to analyzing human behavior known as *rational analysis*),
- Cognitive, perception, and motor operations have their own sets of constraints (an approach known as *bounded cognition*).

Suboptimal human performance results when cognitive systems violate the design assumptions of the environment they evolved to operate within. Optical illusions may be produced because of assumptions made about the environment,<sup>626</sup> during the processing of visual inputs that; the assumptions are beneficial because they simplify the processing of ecologically common inputs. For instance, the human visual system assumes light shines from above, because it has evolved in an environment where this is generally true. A consequence of the assumption of light shining from above in Figure 2.2, is that the top row appears as mounds while the lower row appears as depressions.

Optical illusions are accepted as curious anomalies of the eye/brain system; there is no rush to conclude that human eyesight is faulty. Failures of the cognitive system to produce answers in agreement with mathematical principles, chosen because they appeal to those doing the

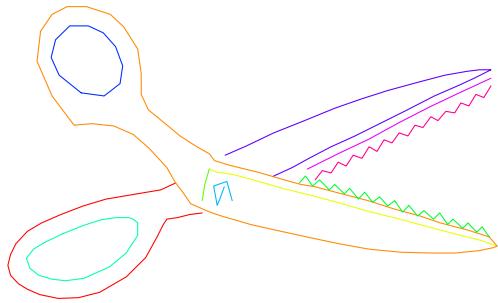


Figure 2.1: Unless cognition and the environment in which it operates closely mesh together, no problems are solved; the blades of a pair of scissors need to closely mesh for cutting to occur. [code](#)

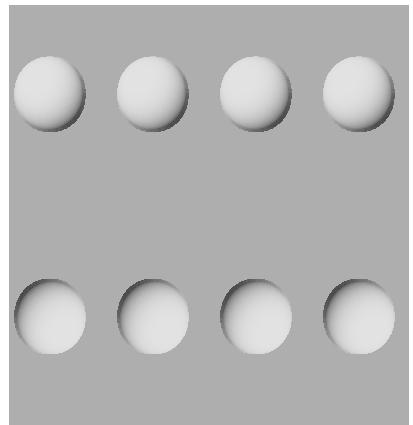


Figure 2.2: The assumption of light shining from above creates the appearance of bumps and pits. [code](#)

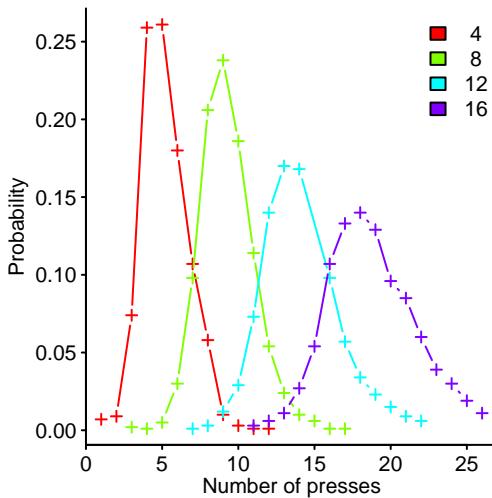


Figure 2.3: Probability that rat N1 will press a lever a given number of times before pressing a second lever to obtain food, when the target count is 4, 8, 12 and 16. Data extracted from Mechner.<sup>929</sup> code

chose, is a signal that the cognitive system has been not been tuned by the environment in which it operates to produce answers compatible with the selected mathematical principles.

This book is written from the viewpoint that the techniques used by people to produce software systems should to be fitted around the characteristics of the computing platform in our head (the view that developers should aspire to be omnipotent logicians is driven by human self-image and is a counter-productive mindset to hold).<sup>i</sup> Builders of bridges do not bemoan the lack of unbreakable materials available to them, they have learned how to work within the limitations of the materials available.

Evolutionary psychology<sup>93, 96</sup> is an approach to psychology which uses knowledge and principles from evolutionary biology to help understand the operation of the human mind. Of course physical implementation details, the biology of the brain,<sup>724</sup> also have an impact on psychological performance.

The fact that cognitive abilities have benefits that outweigh their costs means they are to be found in many creatures,<sup>118</sup> e.g., use of numbers by monkeys<sup>586</sup> and syntax by birds.<sup>1329</sup> A study by Mechner<sup>929</sup> rewarded rats with food if they pressed a lever  $N$  times (with  $N$  taking one of the values 4, 8, 12 or 16), followed by pressing a second lever. Figure 2.3 suggests that rat N1 makes use of an approximate number system.

Table 2.1 shows a division of human time scales by the kind of action that fits within each interval.

Scale (sec)	Time Units	System	World (theory)
10000000	months		Social Band
1000000			
100000			
10000	hours	Task	Rational Band
1000		Task	
100		minutes	
10	10 sec	Task	Cognitive Band
1		Operations	
0.1		Deliberate act	
0.01	10 msec	Unit task	Biological Band
0.001		Operations	
0.0001		Deliberate act	
		Neural circuit	
		Neuron	
		Organelle	

Table 2.1: Time scales of human action. Based on Newell.<sup>1003</sup>

Does the brain contain a collection of modules (each handling particular functionality) or a general purpose processor? This question is the nature vs. nurture debate argument, rephrased using implementation details, i.e., a collection of modules pre-specified by nature or a general purpose processor can be influenced by nurture. The term *modularity of mind* refers to a model of the brain<sup>451</sup> containing a general purpose processor attached to special purpose modules that handle perceptual processes (e.g., hearing and sight); The term *massive modularity hypothesis* refers to a model<sup>292</sup> that only contains modules.

Consciousness is the tip of the iceberg, most of what goes on in the mind is handled by the unconscious.<sup>308, 1438</sup> Problems that are experienced as easy to solve may actually require very complicated neural circuitry and be very difficult for computers to solve...

The only software engineering activities that could be said to be natural, in that there are prewired biological structures in the brain, involve social activities. The exactitude needed for coding is at odds with the fast and frugal approach of our unconscious mind,<sup>497</sup> whose decisions our conscious mind later does its best to justify.<sup>1438</sup> Reading and writing are not natural in the sense that specialist brain structures have evolved to perform these activities; it is the brain's generic ability to learn that enables this skill to be developed.

What are the likely differences in cognitive performance between males and females? A study by Strand, Deary and Smith<sup>1321</sup> analyzed Cognitive Abilities Test (CAT) scores from

<sup>i</sup> An analysis of the operation of human engineering suggests that attempting to modify our existing cognitive systems is a bad idea,<sup>160</sup> e.g., it is better to rewrite spaghetti code than try to patch it.

over 320,000 school pupils in the UK. Figure 2.4 provides a possible explanation for the prevalence of males at the very competent/incompetent ends of the scale and shows that women outnumber men in the middle competency band.

A study by Jørgensen and Grimstad<sup>715</sup> asked subjects from three Asian and three east European countries to estimate the number of lines of code they wrote per hour and the effort needed to implement a specified project (both as part of a project investigating cognitive biases). A regression model built using the results included both country and gender as significant predictors (see [developers/estimation-biases.R](#)).

While much has been written on how society exploits women, relatively little has been written on how society exploits men.<sup>107</sup> There are far fewer women than men directly involved in software engineering.<sup>ii</sup>

### 2.1.1 Models of human cognitive performance

Models of human cognitive performance are based on the results of experiments using samples drawn almost entirely from Western, Educated, Industrialized, Rich and Democratic (WEIRD) societies.<sup>603</sup> While this is a problem for those seeking to uncover the characteristics of humans in general, it is not a problem for software engineering because those involved have often had a WEIRD society education.

Characteristics of WEIRD people that appear to differ from the general population include:

- WEIRD people are willing to think about and make inferences about abstract situations, without the need for having had direct experience, while studies<sup>871</sup> of non-WEIRD people have found they are unwilling to discuss situations where they don't have direct experience,
- when mapping numbers onto space WEIRD people have been found to use a linear scale for mapping values between one and ten, while studies of non-WEIRD people have found they often use a logarithmic scale,<sup>337</sup>
- WEIRD people have been found to have complex, but naive, models of the mechanisms that generate every day random events they observe.<sup>35</sup>

Much of the research carried out in cognitive psychology draws its samples from people between the ages of 18 and 21, studying some form of psychology degree. There has been discussion<sup>92</sup> on the extent to which these results can be extended to the general populace, however, results obtained by sampling from this subpopulation are likely to be good enough for dealing with software engineering issues.

The reasons why students are not appropriate subjects to use in software engineering experiments whose results are intended to be applied to professional software developers are discussed in Chapter 13.

Several executable models of the operation of human cognitive processes have been created. The ACT-R model<sup>38</sup> has been applied to a wide range of problems, including learning, the visual interface, perception and action, cognitive arithmetic, and various deduction tasks.

Studies<sup>460</sup> have found poor correlation between an individual's estimate of their own cognitive ability and measurements of their ability.

### 2.1.2 Embodied cognition

Embodied cognition is the theory that many, if not all, aspects of a person's cognitive processing are dependent on, or shaped by, sensory, motor, and emotional processes that are grounded in the features of their physical body.<sup>503</sup>

A study by Presson and Montello<sup>1120</sup> asked two groups of subjects to memorize the locations of objects in a room. Both groups were then blindfolded and asked to point to various objects;

<sup>ii</sup> When your author started working in software development, if there was a woman working on a team the chances were that she would be at the very competent end of the scale (male/female ratio back then was what, 10/1?). These days, based on my limited experience, women are less likely to be as competent as they once were but still a lot less likely, than men, to be completely incompetent; is the small number of incompetence women caused by a failure of equal opportunity regulations or because the underlying population is small?

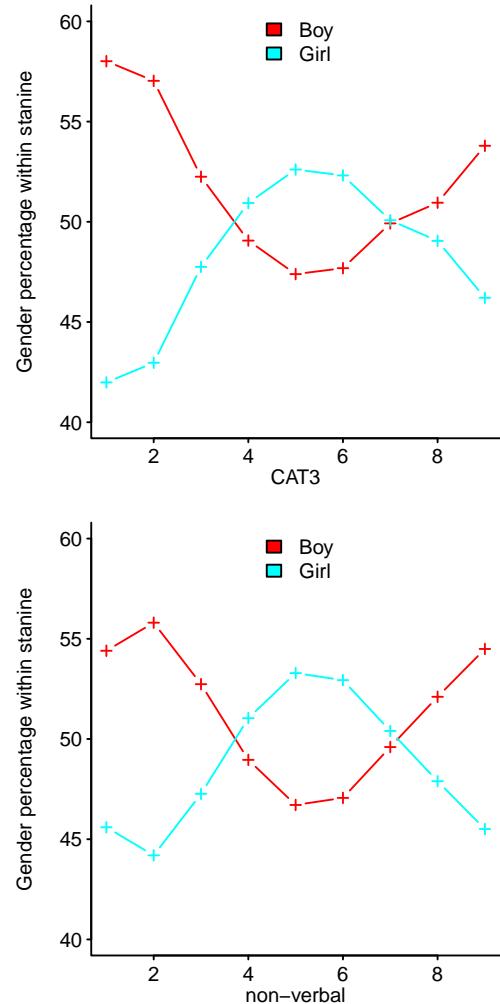


Figure 2.4: Boy/girl (aged 11-12 years) verbal reasoning, quantitative reasoning, non-verbal reasoning and mean CAT score over the three tests; each stanine band is 0.5 standard deviations wide. Data from Strand et al.<sup>1321</sup> [code](#)

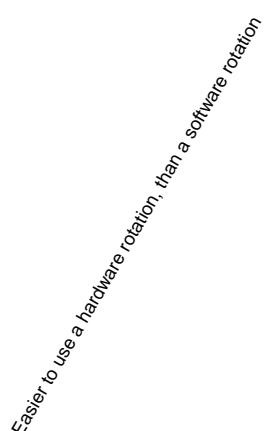


Figure 2.5: Rotate text in the real world, by tilting the head, or in the mind? [code](#)

their performance was found to be reasonably fast and accurate. Subjects in one group were then asked to imagine rotating themselves 90°, they were then asked to point to various objects. Their performance was found to be much slower and less accurate. Subjects in the other group were asked to actually rotate 90°; while still blindfolded, they were then asked to point to various objects. The performance of these subjects was found to be as good as before they rotated. These results suggest that mentally keeping track of the locations of objects, a task that might be thought to be cognitive and divorced from the body, is in fact strongly affected by body position.

Tetris players have been found to prefer rotating an item on screen, as it descends, rather than mentally perform the rotation.<sup>761</sup>

A study by Shepard and Metzler<sup>1243</sup> showed subjects pairs of figures and asked if they were the same. Some pairs were different, while others were the same, but had been rotated relative to each other. The results showed a linear relationship between the angle of rotation (needed to verify that two objects were the same) and the time taken to make a matching comparison. Readers might like to try rotating, in their mind, the pairs of images in Figure 2.6 to find out if they are the same.

A related experiment by Kosslyn<sup>785</sup> showed subjects various pictures and asked questions about them. One picture was of a boat and subjects were asked a question about the front of the boat and then asked a question about the rear of the boat. The response time, when the question shifted from the front to the rear of the boat, was longer than when the question shifted from one about portholes to one about the rear. It was as if subjects had to scan their image of the boat from one place to another to answer the questions.

Many WEIRD people use a mental left-to-right spatial orientation for the number line. This mental orientation has an embodiment in the *SNARC effect* (spatial numerical association of response codes). Studies<sup>335, 1022</sup> show subjects single digit values and ask them to make an odd/even decision by pressing the left/right response button with the left/right hand; when using the right-hand response time decreases as the value increases (i.e., the value moves from left to right along the number line) and when using the left-hand response time decreases as the value decreases. The effect persists when arms are closed, such that opposite hands are used for button pressing. Figure 2.7 shows the error rate for the left/right hand.

## 2.2 Motivation

Motivation to complete a task can have a powerful effect on an individual's priorities, causing them to change in ways that they would not choose in more relaxed circumstances. A study by Darley and Batson<sup>313</sup> asked subjects (theological seminary students) to walk across campus to deliver a sermon. Some subjects were told that they were late and the audience was waiting, the remainder were not told this. Their journey took them past a victim moaning for help in a doorway. Only 10% of subjects who thought they were late stopped to help the victim, while 63% of the other subjects stopped to help. These results do not match the generally perceived behavior pattern of theological seminary students.

Hedonism is an approach to life that aims to maximise personal pleasure and happiness. Many theories of motivation take as their basis that people intrinsically seek pleasure and avoid pain, i.e., they are driven by *hedonic motivation*.

People involved in work that requires creativity can choose to include personal preferences and desires in their decision-making process, i.e., they are subject to hedonic motivation. To date, most of the research on hedonic motivation research has involved studies of consumer behavior...

One widely cited theory that has been gathering experimental support is *Regulatory focus theory*. Regulatory focus is based around the idea that people's different approaches to pleasure and pain influences the approach they take towards achieving an end state (or an end goal). The theory contains two end states, one concerned with aspirations and accomplishments (a *promotion focus*), and the other concerned with attainment of responsibilities and safety (a *prevention focus*).

A promotion focus is sensitive to presence and positive outcomes, seeks to insure hits and insure against errors of omission. A prevention focus is sensitive to absence and negative outcomes, seeks to insure against correct rejections and insure against errors of commission.

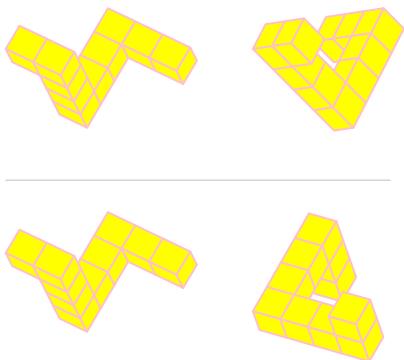


Figure 2.6: Two objects paired with another object that may be a rotated version. Based on Shepard et al.<sup>1243</sup> [code](#)

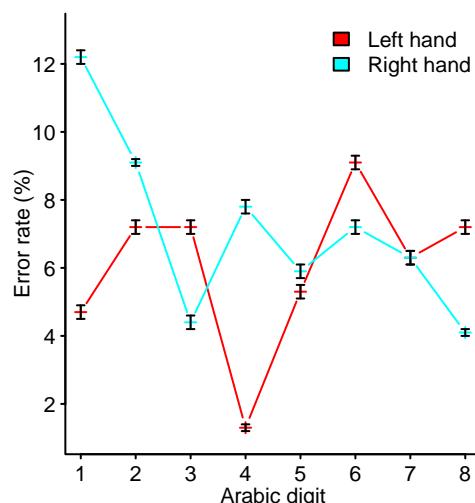


Figure 2.7: Error rate, with standard error, for the left/right hand in a study of the SNARC effect. Data from Nuerk et al.<sup>1022</sup> [code](#)

People are able to exercise some degree of executive control over the priorities given to cognitive processes, e.g., deciding on speed/accuracy trade-offs. Studies<sup>455</sup> have found that subjects with a promotion focus will prefer to trade-off accuracy for speed of performance and those with a prevention focus will trade-off speed for improved accuracy.

The concept of *Regulatory fit*<sup>614</sup> has been used to explain why people engage more strongly with some activities and "feel right" about it (because the activity sustains, rather than disrupts, their current motivational orientation or interests).

A study by Luthiger and Jungwirth<sup>873</sup> investigated the importance of fun as a motivation for software development. Your author was unable to build a good model using the data from the survey of 1,330 developers (e.g., treating percentage of spare time devoted to developing software as a measure of fun), see: [developers/LuthigerJungwirth.R](#).

What motivations are developers attempting to satisfy when they read and write source code? Possible aims and motivations include:

- Performing their role in a development project (with an eye on promotion, for the pleasure of doing a good job, or doing a job that pays for other interests),
- Minimizing cognitive effort; for instance, using heuristics rather than acquiring all the necessary information and applying deductive logic,
- maximizing the pleasure they get out of the time spent writing software...
- belief maintenance: studies have found that people interpret evidence in ways that will maintain their existing beliefs...

?

## 2.3 Memory systems

Memory evolved to supply useful, timely information to an organism's decision-making systems.<sup>769</sup> Memory subsystems are scattered about the brain, with each subsystem/location believed to process and store distinct kinds of information. Figure 2.8 illustrates the current model of known long-term memory subsystems and the region of the brain where they are believed to operate.

Declarative memory has two components that process specific instances of information, one handles facts about the world, while the other episodic memory, deals with events (i.e., the capacity to experience an event in the context in which it occurred; it is not known if non-human brains support episodic memory). We are consciously aware of declarative memory, facts and events can be consciously recalled; it is the kind of memory that is referred to in everyday usage as *memory*. Declarative memory is representational, it can be true or false.

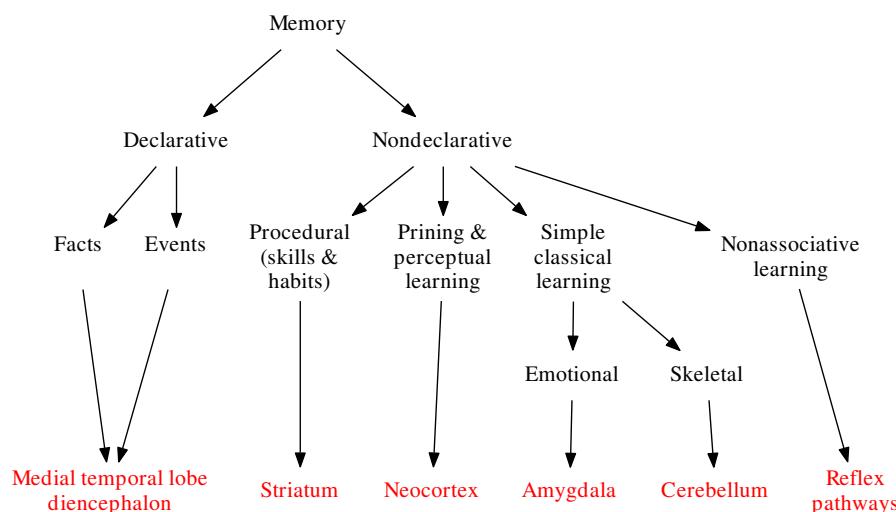


Figure 2.8: Structure of mammalian long-term memory subsystems; brain areas in red. Based on Squire et al.<sup>1292</sup>

Nondeclarative memory (also known as *implicit memory*) extracts information from recurring instances of an experience to form skills (e.g., speaking a language) and habits, simple forms

of conditioning, priming (response to a stimulus is modified by a preceding stimulus;<sup>1153</sup> an advantage in a slowly changing environment where similar events are likely to occur on a regular basis), and perceptual learning (gradual improvement in the detection or discrimination of visual stimuli with practice).

Information in nondeclarative memory is extracted through unconscious performance, e.g., riding a bike; it is an unconscious memory and is not available for conscious recall (information use requires reactivation of the subsystem where the learning originally occurred).

These subsystems operate independently and in parallel, which creates the possibility of conflicting inputs to higher level systems. For instance, a sentence containing the word **blue** may be misinterpreted because information about the word and the color in which it appears, **green**, is returned by different memory subsystems (known as the *Stroop effect*).

A Stroop effect has also been found to occur with lists of numbers. Readers might like to try counting the number of characters occurring in each row in the outside margin. The effort of counting the digit sequences is likely to have been greater and more error prone than for the letter sequences.

Studies<sup>1064</sup> have found that when subjects are asked to enumerate visually presented digits, the amount of Stroop-like interference depends on the arithmetic difference between the magnitude of the digits used and the quantity of those digits displayed. Thus, a short, for instance, list of large numbers is read more quickly and with fewer errors than a short list of small numbers. Alternatively a long list of small numbers (much smaller than the length of the list) is read more quickly and with fewer errors than a long list of numbers where the number has a similar magnitude to the length of the list.

Being able to take advantage of any patterns that occur in an environment is a useful skill.

A study by Reber and Kassin<sup>1149</sup> asked subjects to memorize sets of words. Some words had been generated using a finite state grammar and the rest had not been generated according to the rules of this grammar. There were two groups of subjects, one group was not told about the existence of a pattern in the letter sequences, the other group was told and it was suggested that deducing this pattern could help them to remember the words. The results showed that performance of the group that had not been told about the presence of a pattern almost exactly mirrored that of the group who had been told, on all sets of words (pattern words only, pattern plus non-pattern words, non-pattern words only).

A study by Jones<sup>700</sup> investigated developer beliefs about binary operator precedence. Subjects were shown an expression containing two binary operators, and had to specify the relative precedence of these operators by adding parenthesis to the expression, e.g.,  $a + b | c$ . The more often a pair of binary operators appear together in code, the more often developers have to recall information about their precedence; the hypothesis was that the more often developers have to make a particular precedence decision, when reading code, the more likely they are to know the correct answer. Binary operator usage in code is used as a proxy for developer experience of making binary operator decisions. Figure 2.9 shows how the fraction of correct answers to the relative operator precedence question, with the corresponding percentage occurrence of that pair operator within an expression in C source code.

There has been some research on the interaction between human memory and software development. For instance, Altmann<sup>27</sup> built a computational process model, based on SOAR, and fitted it to 10.5 minutes of programmer activity (debugging within an emacs window); the simulation was used to study the memories, called near-term memory by Altmann, built up while trying to solve a problem.

### 2.3.1 Short term memory

As its name implies, *short term memory* (STM) is the ability to hold information in memory for short periods of time. Short term memory is the popular term for what cognitive psychologists call *working memory*, named after its function rather than the relative duration it can hold information. Early researchers explored its capacity and a paper by Miller<sup>950</sup> introduced the now-famous  $7 \pm 2$  rule. Things have moved on in the 60 years since the publication of his paper<sup>698</sup> (not that Miller ever proposed  $7 \pm 2$  as the capacity of STM; he simply drew attention to the fact that this range of values fitted the results of several experiments).

Readers might like to try measuring their STM capacity using the list of numbers in the outside margin. Any Chinese-speaking readers can try this exercise twice, using the English

and Chinese words for the digits (use of Chinese should enable readers to apparently increase the capacity of STM) in the outside margin. Slowly and steadily read the digits in a row, out loud. At the end of each row, close your eyes and try to repeat the sequence of digits in the same order. If you make a mistake, go on to the next row. The point at which you cannot correctly remember the digits in any two rows of a given length indicates your capacity limit—the number of digits in the previous rows.

The performance impact of reduced working memory capacity can be shown by having people perform two tasks simultaneously. A study by Baddeley<sup>79</sup> measured the time taken to solve a simple reasoning task (e.g.,  $B \rightarrow A$ : “A follows B” True or False?) while remembering a sequence of digits (number of digits is known as the *digit load*). Figure 2.10 shows response time (left axis) and percentage of incorrect answers (right axis) for various digit loads.

Measuring memory capacity using sequences of digits relies on a variety of assumptions, such as assuming all items consume the same amount of memory resources (e.g., digits and letters are interchangeable), that relative item ordering is implicitly included in the measurement and that individual concepts are the unit of storage. Subsequent studies resulted in completely different models of STM being created. What the preceding exercise measured was the amount of *sound* that could be held STM. The sound used to represent digits in Chinese is shorter than in English and using Chinese should enable readers to maintain information on more digits (average 9.9<sup>638</sup>) using the same amount of sound storage. A reader using a language for which the sound of the digits is longer would be able to maintain information on fewer digits, e.g., average 5.8 in Welsh<sup>385</sup> and the average for English is 6.6.

The original observation of a  $7 \pm 2$  capacity limit derives from the number of English digits spoken in two seconds of sound (people speak at different speeds and this is one source of variation included in the  $\pm 2$ ). The two seconds estimate is based on the requirement to remember items and their relative order; the contents of STM do not get erased after two seconds, this limit is the point at which degradation of its contents start to become noticeable.<sup>973</sup> If recall of item order is not relevant, then the limit increases because loss of this information is not relevant.

Studies<sup>1027</sup> involving multiple tasks have been used to distinguish the roles played by various components of working memory (e.g., storage, processing, supervision, and coordination). Figure 2.11 shows the components believed to make up working memory, each with its own independent temporary storage areas, each holding and using information in different ways.

The central executive is assumed to be the system that handles attention, controlling the phonological loop, the visuo-spatial sketch pad, and the interface to long-term memory. The central executive needs to remember information while performing tasks such as text comprehension and problem solving. It has been suggested that the focus of attention is capacity-limited, but that the other temporary storage areas are time-limited (without attention to rehearse them, they fade away)<sup>297</sup>...

Visual information held in the visuo-spatial sketch pad decays very rapidly. Experiments have shown that people can recall four or five items immediately after they are presented with visual information, but that this recall rate drops very quickly after a few seconds. From the source code reading point of view, the visuo-spatial sketch pad is only operative for the source code currently being looked at.

Mental arithmetic provides an example of how different components of working memory can be combined to solve a problem that is difficult to achieve using just one component; multiply 23 by 15 without looking at this page. All information has to be held in short term memory and the central executive. Now perform another multiplication, but this time look at the two numbers being multiplied (see outer margin for values) while performing the multiplication.

Looking at the numbers reduces the load on working memory. Multiplying while being able to look at the numbers being multiplied seems to require less cognitive effort.

Table 2.2 contains lists of words; those at the top of the table contain a single syllable, those at the bottom multiple syllables. Readers should have no problems remembering a sequence of five single-syllable words, a sequence of five multi-syllable words should prove more difficult. As before, read each word slowly out loud.

It has been found that fast talkers have better short-term memory. The connection is the phonological loop. Short-term memory is not limited by the number of items that can be held, but the length of sound that can be stored (about two seconds<sup>81</sup>). Faster talkers can represent more information in that two seconds than those who do not talk as fast.

8704
2193
3172
57301
02943
73619
659420
402586
542173
6849173
7931684
3617458
27631508
81042963
07239861
578149306
293486701
721540683
5762083941
4093067215
9261835740

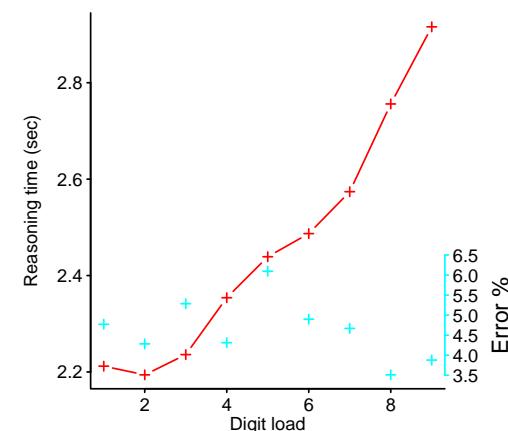


Figure 2.10: Response time (left axis) and error percent (right axis) on reasoning task with given number of digits held in memory. Data extracted from Baddeley.<sup>79</sup> code

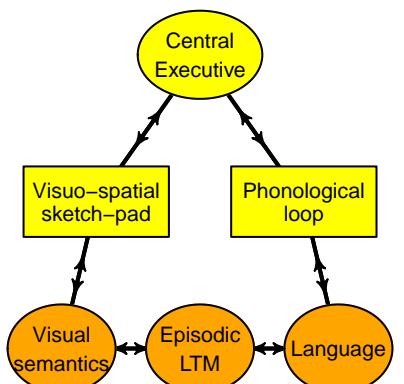


Figure 2.11: Major components of working memory: working memory in yellow, long-term memory in orange. Based on Baddeley.<sup>80</sup> code

List 1	List 2	List 3	List 4	List 5
one	cat	card	harm	add
bank	lift	list	bank	mark
sit	able	inch	view	bar
kind	held	act	fact	few
look	mean	what	time	sum
ability	basically	encountered	laboratory	commitment
particular	yesterday	government	acceptable	minority
mathematical	department	financial	university	battery
categorize	satisfied	absolutely	meaningful	opportunity
inadequate	beautiful	together	carefully	accidental

Table 2.2: Words with either one or more than one syllable (and thus varying in the length of time taken to speak).

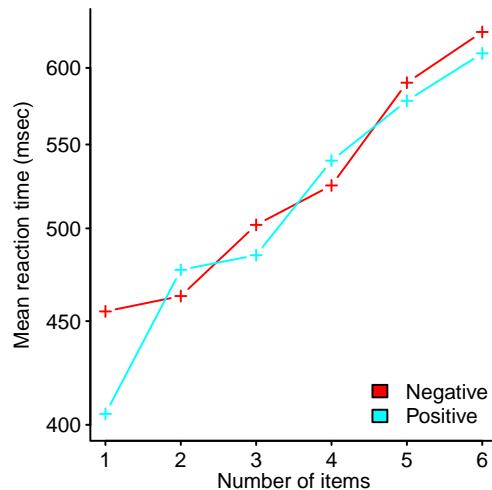


Figure 2.12: Yes/no response time (in milliseconds) as a function of the number of digits held in memory. Data extracted from Sternberg.<sup>1312</sup> [code](#)

An analogy between *phonological loop* and a loop of tape in a tape recorder, suggests the possibility that it might only be possible to extract information as it goes past a *read-out point*. A study by Sternberg<sup>1312</sup> asked subjects to hold a sequence of digits in memory, e.g., 4185 and measured the time taken to respond yes/no on whether a particular digit was in this sequence. Figure 2.12 shows that as the number of digits increased, the time taken for subjects to respond increases. The other result was that response time was not affected by whether the answer was yes or no. It might be expected that a yes answer would enable searching to terminate, but the behavior found suggests that all digits were always being compared. Different kinds of information has different search response times.<sup>222</sup>

Extrapolating the results from studies based on the use of natural language,<sup>316</sup> to the use of computer languages, need to take into account that reader performance has been found to differ between words (character sequences having a recognized use in the reader's human language) and nonwords, e.g., naming latency is lower for words,<sup>1437</sup> more words can be held in short term memory<sup>650</sup> ( $\text{word\_span} = 2.4 + 2.05 \times \text{speech\_rate}$ , and  $\text{nonword\_span} = 0.7 + 2.27 \times \text{speech\_rate}$ ).

The ability to comprehend syntactically complex sentences is correlated with working memory capacity.<sup>755</sup> A study by Miller and Isard<sup>951</sup> investigated subjects' ability to memorize sentences that varied in their degree of embedding. The following sentences have increasing amounts of embedding (Figure 2.13 shows the parse-tree of two of them):

She liked the man that visited the jeweller that made the ring that won the prize that was given at the fair.

The man that she liked visited the jeweller that made the ring that won the prize that was given at the fair.

The jeweller that the man that she liked visited made the ring that won the prize that was given at the fair.

The ring that the jeweller that the man that she liked visited made won the prize that was given at the fair.

The prize that the ring that the jeweller that the man that she liked visited made won was given at the fair.

Subjects' ability to correctly recall wording decreased as the amount of embedding increased, although performance did improve with practice. People have significant comprehension difficulties when the degree of embedding in a sentence exceeds two.<sup>150</sup>

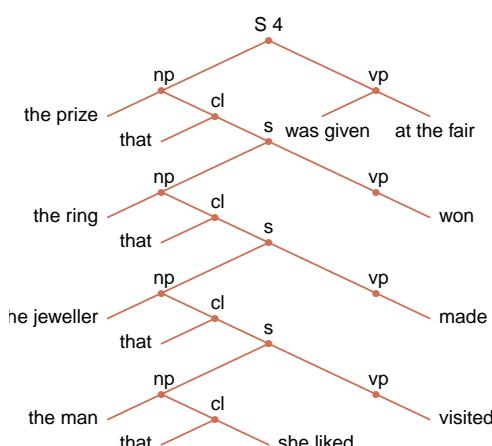
Human language has a grammatical structure that enables it to be parsed serially (e.g., as it is spoken).<sup>1086</sup> A consequence of this expected characteristic of sentences is that so called *garden path* sentences, where one or more words at the end of a sentence changes the parsing of words read earlier, generate confusion (requiring conscious effort to reason about what has been said):

The old train their dogs.

The patient persuaded the doctor that he was having trouble with to leave.

While Ron was sewing the sock fell on the floor.

Joe put the candy in the jar into my mouth.



The horse raced past the barn fell.

### 2.3.2 Episodic memory

Episodic memory is memory for personally experienced events that are remembered as such, i.e., the ability to recollect specific events or episodes in our lives. When the remembered events occurred sometime ago, the term *autobiographical memory* might be used.

What impact does the passage of time have on episodic memories?

A study by Altmann, Trafton and Hambrick<sup>29</sup> investigated the effects of interruption on a task involving seven steps. Subjects performed the same task 37 times and were interrupted at random intervals during the 30-50 minutes it took to complete the session. Interruptions required subjects to perform a simple typing task that took, on average, 2.8, 13, 22 and 32 seconds. Figure 2.14 shows the percentage of sequencing errors made immediately after an interruption and under normal working conditions (a sequence error occurs when an incorrect step is performed, e.g., step 5 is performed again after performing step 5, when step 6 should have been performed; the offset on the x-axis is the difference between the step performed and the one that should have been performed; the sequence error rate, as a percentage of the total number of tasks performed at each interruption interval, was 2.4, 3.6, 4.6 and 5.1%). The lines are predictions made by a model fitted to the data.

### 2.3.3 Forgetting

People are unhappy when they forget things, however not forgetting may be worse.<sup>1016</sup> The Russian mnemonist Shereshevskii found that his ability to remember everything cluttered up his mind.<sup>872</sup> Having many similar, not recently used, pieces of information matching during a memory search would be counterproductive; forgetting is a useful adaptation.<sup>1212</sup> For instance, a driver returning to a car wants to know where it was last parked, not the location of all previous parking locations. It has been proposed that human memory is optimized for information retrieval based on the statistical properties of likely need for the information,<sup>39</sup> in people's everyday lives; Burrell investigated the pattern of book borrowings in several libraries; which were also having items added to their stock.<sup>203</sup> The rate at which the mind forgets seems to mirror the way that information tends to lose its utility in the real world over time.

Forgetting, like learning, follows a power law<sup>1184</sup> (the results of some studies are also well fitted by an exponential equation). The general relationship between the retention of information,  $R$ , and the time,  $T$ , since the last access has the form  $R = aD^{-b}$  (where  $a$  and  $b$  are constants). It is known as the *power law of forgetting*.

emailed for data...?

### 2.3.4 Recognition and recall of information

Kinds of information retrieval from memory...

- recognition: studies<sup>782</sup> have found that people can often make a reasonably accurate judgement about whether they know a piece of information or not, even if they are unable to recall that information at a particular instant; the so-called *feeling of knowing* is a good predictor of subsequent recall of information,
- recall: ... models of information recall<sup>191</sup> over the human timescales, short and long...

The environment in which information was learned can have an impact on recall performance. A study by Godden and Baddeley<sup>505</sup> investigated subjects' recall of memorized words in two different environments. Subjects were divers and learned a list of spoken words either while submerged underwater wearing scuba apparatus or while sitting at a table on dry land. The results showed that subjects recall performance was significantly better when performed in the environment in which the word list was learned.

When asked to retrieve members of a category, people tend to produce a list of semantically related items, before switching to list another cluster of semantically related items and so on.

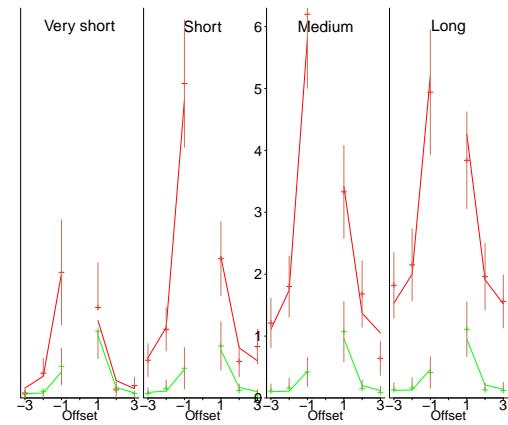


Figure 2.14: Sequencing errors (as percentage) after interruptions of various length (red), including 95% confidence intervals, normal sequence error rate in green; lines are fitted model predictions. Data from Altmann et al.<sup>29</sup> code

This pattern of retrieval is similar to optimal food foraging strategies,<sup>618</sup> however the same behavior would also emerge from a random walk on a semantic network built from human word-association data.<sup>1</sup>

Chunking is a technique commonly used by people to help them remember information. A chunk is a small set of items ( $4\pm1$  is seen in many studies) having a common, strong, association with each other (and a much weaker one to items in other chunks). For instance, Wickelgren<sup>1448</sup> found that people's recall of telephone numbers is optimal if numbers are grouped into chunks of three digits. An example from random-letter sequences is **fbi****cbs****bmir**s. The trigrams (**fbi**, **cbs**, **ibm**, **irs**) within this sequence of 12 letters are well-known acronyms. A person who notices this association can use it to aid recall. Several theoretical analyses of memory organizations have shown that chunking of items improves search efficiency (optimal chunk size 3–4,<sup>356</sup> number items at which chunking becomes more efficient than a single list, 5–7<sup>880</sup>).

A study by Klahr, Chase, and Lovelace<sup>766</sup> investigated how subjects stored letters of the alphabet in memory. Through a series of time-to-respond measurements, where subjects were asked to name the letter that appeared immediately before or after the presented probe letter, they proposed the alphabet-storage structure shown in Figure 2.15.

### 2.3.4.1 Serial order information

A fundamental requirement for most, if not all, behaviors is the ability to process serial order information. Sequential processing is central to verbal behaviors (e.g., speech perception and generation) and other behaviors ranging from motor control to planning, and goal-directed action. How behaviors are sequenced is one of the most important problems in psychology.

Studies<sup>656</sup> have consistently found a variety of patterns in recall of serial information, such as a list of recently remembered items; patterns include the following:

- better recall performance for items at the start (the *primacy effect*) and end (the *recency effect*) of a list (known as the *serial position effect*,<sup>981</sup> see Figure 2.16),
- recall of a short list tends to start with the first item and progress in order through the list, while for a long list people are more likely to start with one of the last four items;<sup>1427</sup> when prompted by an entry on the list people are most likely to recall the item following it<sup>643</sup> (see [developers/misc/HKJEP99.R](#)).

A study by Pennington<sup>1075</sup> found that developers responded slightly faster to questions about a source code statement when its immediately preceding statement made use of closely related variables...

## 2.4 Learning and experience

People have the ability to implicitly and explicitly learn and in human performance on many tasks improves with practice, e.g., time to deciding whether a character sequence is an English word or not<sup>746</sup>; many studies have fitted a power law to practice performance measurements (the term *power law of learning* is often used). If chunking is assumed to play a part in learning, a power law is a natural consequence;<sup>1004</sup> the equation has the form:

$$RT = a + bN^{-c}$$

where  $RT$  is the response time;  $N$  is the number of times the task has been performed; and  $a$ ,  $b$ , and  $c$  are constants.

There are also good theoretical reasons for expecting the measurements to be fitted by an equation having an exponential form and such as equation has been fitted to many data sets;<sup>595</sup> the equation has the form:

$$RT = a + be^{-cN}$$

Implicit learning occurs when people perform a task containing information that is not explicitly obvious to those performing it. A study by Reber and Kassin<sup>1149</sup> compared implicit and explicit pattern detection. Subjects were asked to memorize sets of words, with the words in some sets containing letters generated using a finite state grammar. One group of

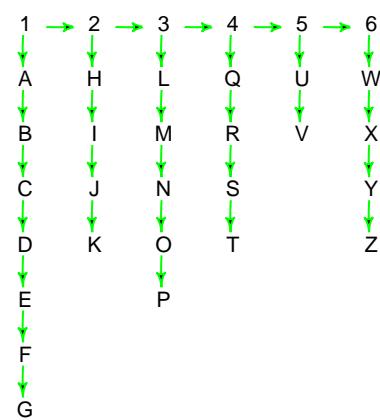


Figure 2.15: Semantic memory representation of alphabetic letters (the numbers listed along the top are place markers and are not stored in subject memory). Readers may recognize the structure of a nursery rhyme in the letter sequences. Derived from Klahr.<sup>766</sup> [code](#)

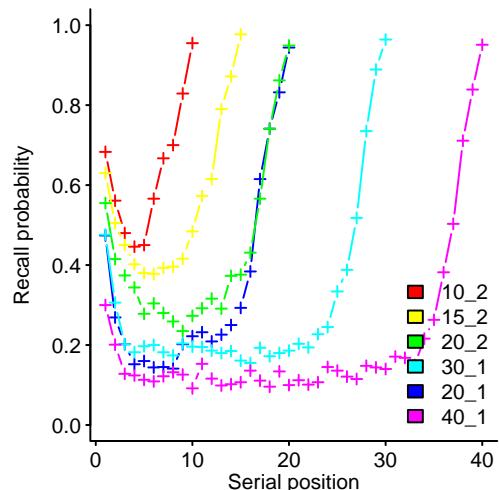


Figure 2.16: Probability of correct recall of words by serial presentation order (each word visible for 1 or 2 seconds, last digit in legend). Data extracted from Murdoch.<sup>981</sup> [code](#)

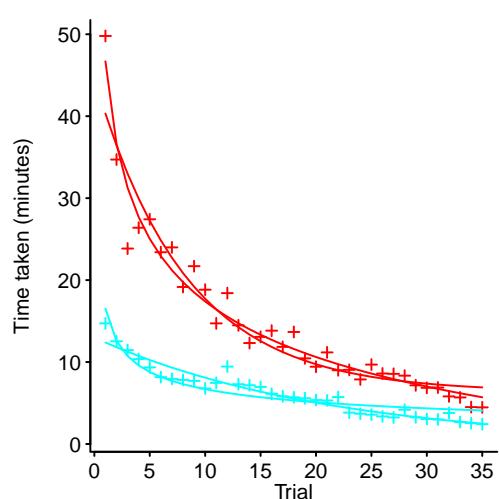


Figure 2.17: Time taken to solve the same jig-saw puzzle 35 times, followed by a two-week interval and then another 35 times, with power law and exponential fits. Data extracted from Alteneder.<sup>26</sup> [code](#)

subjects thought they were just taking part in a purely memory-based experiment, while the second group were told of the existence of a letter sequence pattern in some words and that it would help their performance if they could deduce this pattern. The performance of the two groups on the different sets of words (i.e., pattern words only, pattern plus non-pattern words, non-pattern words only) matched each other. Without being told to do so, subjects had used patterns in the words to help perform the memorization task.

Explicit learning occurs when the task contains obvious patterns that can be remembered and used on subsequent performances. A study by Alteneder<sup>26</sup> recorded the time taken by the author to solve the same jig-saw puzzle 35 times (over a four-day period). After two weeks, the same puzzle was again solved 35 times. The exponent of the fitted power law, for the first series, is -0.5; Figure 2.17 show both a fitted power law and exponential.

For simple problems, learning can result in the solution being committed to memory; performance is then driven by reaction-time, i.e., the time needed to recall and give the response. Logan<sup>860</sup> provides an analysis of subject performance on these kinds of problems.

The source code for an application does not need to be rewritten every time somebody wants a copy; it is unusual for the same developer to be asked to implement exactly the same application again. Having a developer reimplemented the same application many times provides some insight into the underlying

A study by Lui and Chan<sup>867</sup> asked 24 developers to implement the same application four times; 16 developers worked in pairs (i.e., eight pair programming teams) and eight worked solo. Before starting to code, the subjects took a test involving 50 questions from a computer aptitude test; subjects were ranked by number of correct answers and pairs selected such that both members were adjacent in the ranking.

Learning occurs every-time the application is written and forgetting occurs during the period between implementations (each implementation occurred on a separate weekend, with subjects working during the week) . Each subject has existing knowledge and skill, which means everybody starts the experiment at a different point on the learning curve. In the following analysis the test score is used as a proxy for this each subject's initial point on the learning curve.

Figure 2.18 shows the completion times, for each round of implementation, for solo and pairs, along with the predictions from a mixed-effects model fitted using the following equation:

$$\text{Completion\_time} = a \cdot (b \cdot \text{Test\_score} + \text{Round})^c$$

where: *Completion\_time* is time to complete an implementation of the application, *Test\_score* the test score and *Round* is the number of times the application has been implemented; *a*, *b* and *c* are constants chosen by the model building process. While the parameters of this equation can be fitted to better than 0.05 significance, predictions are in poor agreement with actual values, suggesting that additional factors make an important contribution.

When starting work in a new environment, performance can noticeable improve as a person learns how the environment operates. A study by Brooks<sup>189</sup> measured the performance of an experienced developer, writing and debugging 23 short programs (mean length 24 lines). The time taken to debug each program improved as more programs were implemented (see [developers/a013582.R](#)).

Developers sometimes work in several languages on a regular basis. The extent to which learning applies across languages will have some dependency on the easy with which implementation details are applicable across the languages used.

A study by Zislis<sup>1501</sup> measured the time taken (in minutes) by himself to implement 12 algorithms, with the implementation performed three times using three languages (APL, PL/1, Fortran) and on the fourth repetition using the same language as on the first implementation. Figure 2.19 shows total implementation time for each algorithm over the four implementations; a fitted mixed-effects model finds some performance difference between the languages used (see figure code for details).

A study by Mockus and Weiss<sup>964</sup> found that the probability of developer introducing a fault into an application, when modifying the software, decreased as the log of the total number of changes made by the developer (i.e., their experience or expertise).

binary operator precedence...

Job advertisements often specify that a minimum number of years of experience is required. Number of years is known not to be a measure of expertise, but it provides some degree of

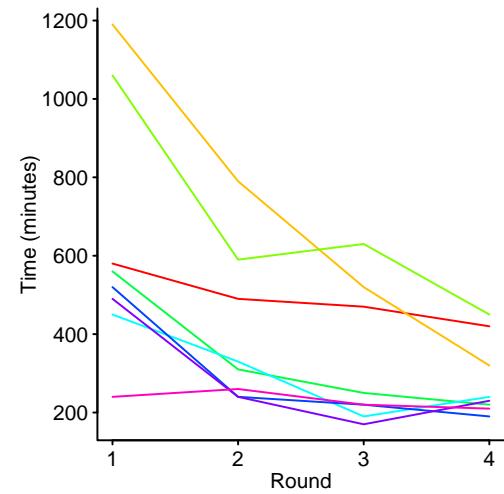


Figure 2.18: Completion times of eight solo developers for each implementation round. Data kindly provided by Lui.<sup>867</sup> [code](#)

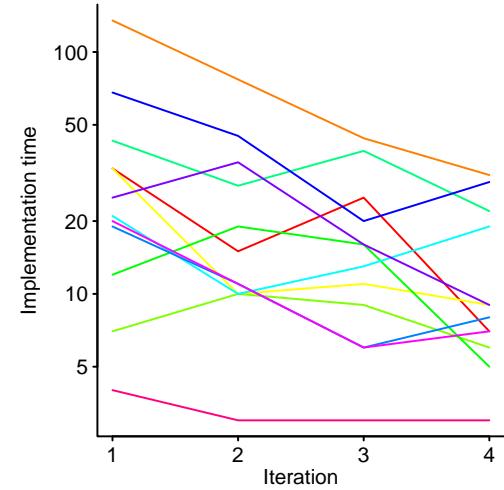


Figure 2.19: Time taken, by the same person, to implement 12 algorithms from the Communications of the ACM (each colored line), with four iteration of the implementation process. Data from Zislis.<sup>1501</sup> [code](#)

comfort that a person has had to deal with many of the problems that might occur within a given domain.

A study by Latorre<sup>819</sup> the effect of developer experience on applying unit-test-driven development. The 24 subjects, classified as junior (one-to-two-years professional experience), intermediate (three-to-five-years experience) or senior (more than six-years experience), were taught unit-test-driven development and the time taken for them to implement eight groups of requirements was measured. The implementation of the first two groups was included as part of the training and familiarization process; the time taken, by each subject, on these two groups was used to normalise the reported results.

Figure 2.20 shows the normalised time taken, by each subject, on successive groups of requirements; color is used to denote subject experience. While there is a lot of variation between subjects, average performance improves with years of experience, i.e., implementation time decreases (a fitted mixed-effects model is included in the code).

The improvement in performance achieved through practice starts to decline when a person stops performing the activity that is generating learning.<sup>676</sup> Clerical task<sup>1250</sup>... emailed for data

To quote Herbert Simon:<sup>1258</sup> “Intuition and judgement—at least good judgement—are simply analyses frozen into habit and into the capacity for rapid response through recognition. . . . Every manager needs also to be able to respond to situations rapidly, a skill that requires the cultivation of intuition and judgement over many years of experience and training.”

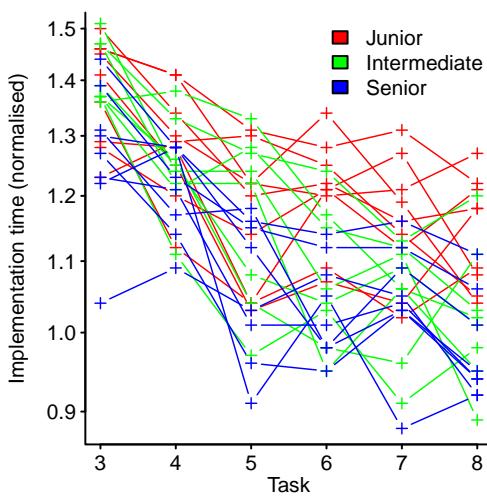


Figure 2.20: Time taken by 24 subjects, classified by years of professional experience, to complete successive tasks. Data from Latorre.<sup>819</sup> code

### 2.4.1 Belief

Knowledge could be defined as belief plus complete conviction and conclusive justification.

- The *foundation approach* argues that beliefs are derived from reasons for these beliefs. A belief is justified if and only if (1) the belief is self-evident and (2) the belief can be derived from the set of other justified beliefs (circularity is not allowed).

This is very costly. in cognitive effort, to operate, e.g., the reasons for beliefs need to be remembered and applied when considering new beliefs. Studies<sup>1177</sup> show that people exhibit a belief preservation effect; they continue to hold beliefs after the original basis for those beliefs no longer holds. The evidence suggests that people use some form of *coherence approach* for creating and maintaining their beliefs.

- the *coherence approach* argues that where beliefs originated is of no concern. Instead, beliefs must be logically coherent with other beliefs (believed by an individual). These beliefs can mutually justify each other and circularity is allowed. A number of different types of coherence have been proposed, Including *deductive coherence* (requires a logically consistent set of beliefs), *probabilistic coherence* (assigns probabilities to beliefs and applies the requirements of mathematical probability to them), *semantic coherence* (based on beliefs that have similar meanings), and *explanatory coherence* (requires that there be a consistent explanatory relationship between beliefs).

**The Belief-Adjustment model** A belief may be based on a single piece of evidence, or it may be based on many pieces of evidence. How is an existing belief modified by the introduction of new evidence? The belief-adjustment model of Hogarth and Einhorn<sup>630</sup> offers an answer to this question; the basic equation for their model is:

$$S_k = S_{k-1} + w_k[s(x_k) - R]$$

where:  $0 \leq S_k \leq 1$  is the degree of belief in some hypothesis or impression after evaluating  $k$  items of evidence;  $S_{k-1}$  is the anchor, or prior opinion (with  $S_0$  denoting the initial belief);  $s(x_k)$  is a subjective evaluation of the  $k$ 'th item of evidence (different people may assign different values for the same evidence,  $x_k$ );  $R$  is the reference point, or background, against which the impact of the  $k$ 'th item of evidence is evaluated;  $0 \leq w_k \leq 1$  is the adjustment weight for the  $k$ 'th item.

When presented with new information people can process the evidence it contains in several ways, including:

- using an *evaluation* process, which encodes new evidence relative to a fixed point—the hypothesis addressed by a belief. If the new evidence supports the hypothesis, a person's belief is increased, and if it is not supported by the hypothesis the belief is decreased. The increase/decrease occurs irrespective of the current state of a person's belief; for this case:  $-1 \leq s(x_k) \leq 1$  and  $R = 0$ , and the belief-adjustment equation simplifies to:

$$S_k = S_{k-1} + w_k s(x_k)$$

An example of an evaluation process might be the belief that the object X always holds a value that is numerically greater than Y.

- using an *estimation* process, which encodes new evidence relative to the current state of a person's beliefs. For this case  $R = S_{k-1}$ , and the belief-adjustment equation simplifies to:

$$S_k = S_{k-1} + w_k [s(x_k) - S_{k-1}]$$

where:  $0 \leq s(x_k) \leq 1$

In this case the degree of belief, in a hypothesis, can be thought of as a moving average. For an estimation process, the order in which evidence is presented can be significant. While reading source code written by somebody else, a developer will form an opinion of the quality of that person's work. The judgement of each code sequence will be based on the readers current opinion (at the time of reading) of the person who wrote it.

The  $s(x_k)$  could represent either the impact of a single piece of evidence (known as *Step-by-Step*, SbS), or several pieces of evidence that have been combined to have a single impact (known as *End-of-Sequence*, EoS).

Hogarth and Einhorn proposed that when people are required to provide an EoS response they use an EoS process when the sequence of items is short and simple. As the sequence gets longer, or more complex, they shift to an SbS process, to keep the peak cognitive load (of processing the evidence) within their capabilities.

**Order effects** Use of an SbS process when  $R = S_{k-1}$  leads to a recency effect.<sup>630</sup> When  $R = 0$ , a recency effect only occurs when there is a mixture of positive and negative evidence (there is no recency effect if the evidence is all positive or all negative).

The use of an EoS process leads to a primacy effect; however, a task may not require a response until all the evidence is seen. If the evidence is complex, or there is a lot of it, people may adopt an SbS process. In this case, the effect seen will match that of an SbS process.

A study by Hogarth and Einhorn<sup>630</sup> investigated order, and response mode, effects in belief updating. Subjects were presented with a variety of scenarios (e.g., a defective stereo speaker thought to have a bad connection or a baseball player whose hitting improved dramatically after a new coaching program). Subjects read an initial description followed by two or more additional items of evidence. The additional evidence was either positive (e.g., “The other players on Sandy's team did not show an unusual increase in their batting average over the last five weeks”) or negative (e.g., “The games in which Sandy showed his improvement were played against the last-place team in the league”). The positive and negative evidence was worded to create either strong or weak forms.

The evidence was presented in three combinations: strong-positive and weak-positive, upper plot in Figure 2.21; strong-negative and weak-negative, middle plot of Figure 2.21; positive-negative and negative-positive, lower plot of Figure 2.21. Subjects were asked, “Now, how likely do you think X caused Y on a scale of 0 to 100?” In some cases, subjects had to respond after seeing each item of evidence: in other cases, subjects had to respond after seeing all the items.

Other studies have duplicated these results, for instance, professional auditors have been shown to display recency effects in their evaluation of the veracity of company accounts.<sup>1061</sup>

## 2.4.2 Category knowledge

Children as young as four have been found to use categorization to direct the inferences they make,<sup>486</sup> and many studies have shown that people have an innate desire to create and use categories (they have also been found to be sensitive to the costs and benefits of using categories<sup>886</sup>). By dividing items they encounter into categories, people reduce the amount of information they need to learn<sup>1108</sup> and to generalize based on prior experience.<sup>1013</sup> Probably

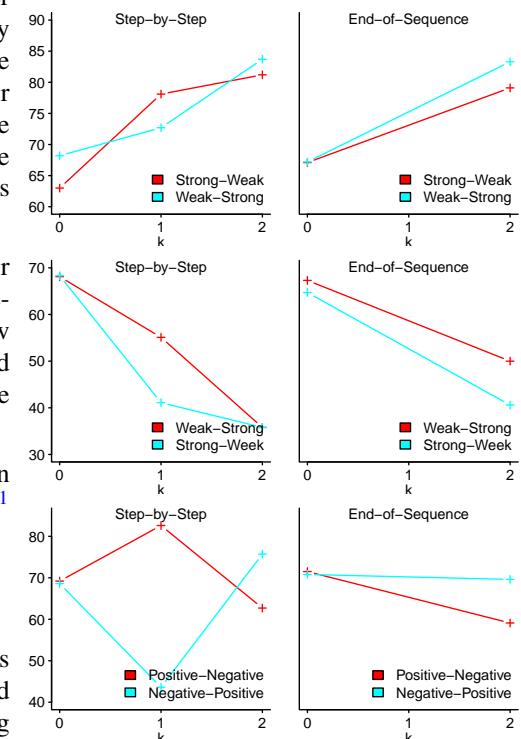


Figure 2.21: Subjects belief response curves for positive weak-strong, negative weak-strong, and positive-negative evidence. Based on Hogarth et al.<sup>630</sup> code

July 31, 2018

information about the characteristics of a newly encountered item is obtained by matching it to one or more known categories and then extracting characteristics common to previously encountered items in those categories. For instance, a flying object with feathers and a beak might be assigned to the category *bird*, which suggests the characteristics of laying eggs and being migratory.

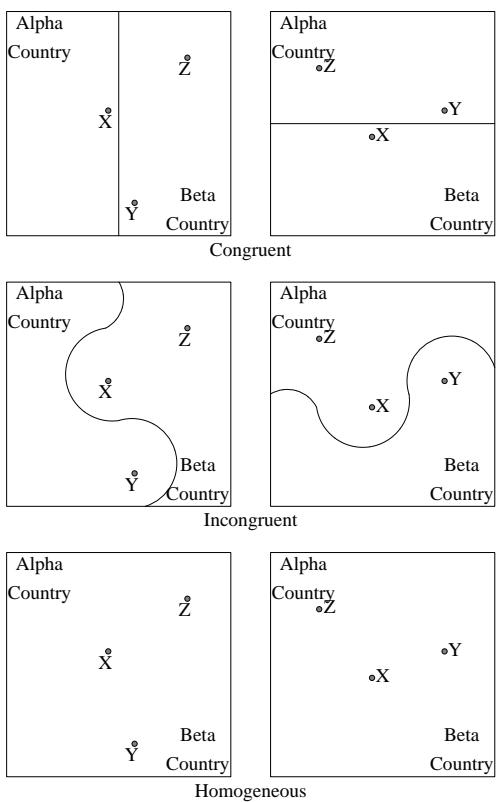


Figure 2.22: Country boundaries distort judgement of relative city locations. Based on Stevens et al.<sup>1313</sup>

Categorization is used to perform inductive reasoning (the derivation of generalized knowledge from specific instances), and also acts as a memory aid (remembering the members of a category). Categories provide a framework from which small amounts of information can be used to infer, seemingly unconnected (to an outsider), useful conclusions.

Several studies have shown that people use around three levels of abstraction in creating hierarchical relationships. Rosch<sup>1173</sup> called the highest level of abstraction the *superordinate-level*—for instance, the general category furniture. The next level down is the *basic-level*; this is the level at which most categorization is carried out—for instance, car, truck, chair, or table. The lowest level is the *subordinate-level*, denoting specific types of objects. For instance, a family car, a removal truck, my favourite armchair, a kitchen table. Rosch found that the basic-level categories had properties not shared by the other two categories; adults spontaneously name objects at this level, it is the abstract level that children acquire first, and category members tend to have similar overall shapes.

When categories have hierarchical structure it is possible for an attribute of a higher-level category to affect the perceived attributes of subordinate categories. This is illustrated in a study by Stevens and Coupe<sup>1313</sup> in which subjects were asked to remember the information contained in a series of maps (see Figure 2.22). They were asked questions such as: “Is X east or west of Y?” and “Is X north or south of Y?” Subjects gave incorrect answers 18% of the time for the congruent maps, but 45% of the time for the incongruent maps (15% for homogeneous). These results were interpreted as information about the relative locations of countries influencing the answer questions about the city locations.

Studies<sup>1269</sup> have found that people do not consistently treat subordinate categories as inheriting the properties of their superordinates, i.e., category inheritance need not be a tree.

How categories should be defined and structured is a long-standing debate within the sciences. Some commonly used category formation techniques, their membership rules and attributes include:

- in the defining-attribute theory members of a category are characterized by a set of defining attributes. Attributes divide objects up into different concepts whose boundaries are well-defined and all members of the concept are equally representative. Also, concepts that are a basic-level of a superordinate-level concept will have all the attributes of that superordinate level; for instance, a sparrow (small, brown) and its superordinate bird (two legs, feathered, lays eggs),
- in the prototype theory categories have a central description, the prototype, that represents the set of attributes of the category. This set of attributes need not be necessary, or sufficient, to determine category membership. The members of a category can be arranged in a typicality gradient, representing the degree to which they represent a typical member of that category. It is also possible for objects to be members of more than one category (e.g., tomatoes as a fruit, or a vegetable),
- in the exemplar-based theory of classification specific instances, or *exemplars*, act as the prototypes against which other members are compared. Objects are grouped, relative to one another, based on some similarity metric. The exemplar-based theory differs from the prototype theory in that specific instances are the norm against which membership is decided. When asked to name particular members of a category, the attributes of the exemplars are used as cues to retrieve other objects having similar attributes,
- in the explanation-based theory of classification there is an explanation for why categories have the members they do. For instance, the biblical classification of food into *clean* and *unclean* is roughly explained by saying that there should be a correlation between type of habitat, biological structure, and form of locomotion; creatures of the sea should have fins, scales, and swim (sharks and eels don't) and creatures of the land should have four legs (ostriches don't).

From a predictive point of view, explanation-based categories suffer from the problem that they may heavily depend on the knowledge and beliefs of the person who formed the cate-

gory; for instance, the set of objects a person would remove from their home if it suddenly caught fire.

Figure 2.23 shows the possible combinations of three, two-valued attributes, color/size/shape; there are eight possibilities. It is possible to create six unique categories by selecting four items from these eight possibilities (see Figure 2.24; there are 70 different ways of taking four things from a choice of eight,  $(8!/(4!4!))$ , and taking symmetry into account reduces the number to unique categories to six).

A study by Shepard, Hovland, and Jenkins<sup>1242</sup> measured subject performance in assigning objects to these six categories. Subject error rate decreased with practice.

A method of calculating the similarity of two objects, proposed by Estes,<sup>399</sup> is in general agreement with the Shepard et al results. A matching similarity coefficient,  $0 \leq t \leq \infty$  is assigned for matching attributes and a nonmatching similarity coefficient,  $0 \leq s_i \leq 1$  (potentially different for each nonmatch), is assigned for each nonmatching attribute. The similarity of two objects is calculated by multiplying the similarity coefficients. When comparing objects within the same category the convention is to use,  $t = 1$  and to give the attributes that differ the same similarity coefficient,  $s$ .

As an example, for simplicity consider the two attributes shape/color in Figure 2.23, giving the four combinations black/white—triangles/squares; assign black triangle and black square to category A, and the white triangle and white square to category B, i.e., category membership is decided by color. The similarity of each of the four possible object combinations to category A and B is listed in Table 2.3. Looking at the top row: Black triangle is compared for similarity to all members of category A ( $1 + s$ , because it does not differ from itself and differs in one attribute from the other member of category A) and all members of category B ( $s + s^2$ , because it differs in one attribute from one member of category B and in two attributes from the other member).

Stimulus	Similarity to A	Similarity to B
Black triangle	$1 + s$	$s + s^2$
Black square	$1 + s$	$s + s^2$
White triangle	$s + s^2$	$1 + s$
White square	$s + s^2$	$1 + s$

Table 2.3: Similarity of a Stimulus object to two categories (category A: black triangle and black square; category B: white triangle and white square).

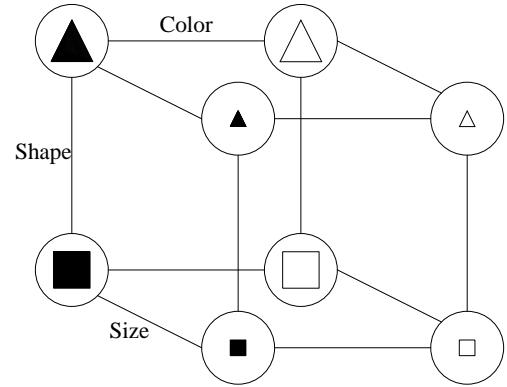


Figure 2.23: Orthogonal representation of shape, color and size stimuli. Based on Shepard.<sup>1242</sup>

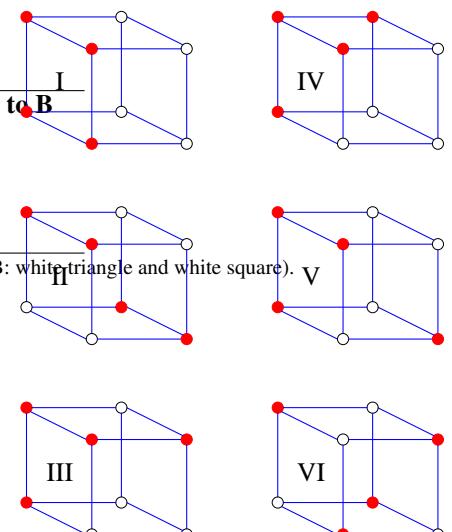


Figure 2.24: The six unique configurations of selecting four times from eight possibilities, i.e., it is not possible to rotate one configuration into another within these six configurations. Based on Shepard.<sup>1242</sup>

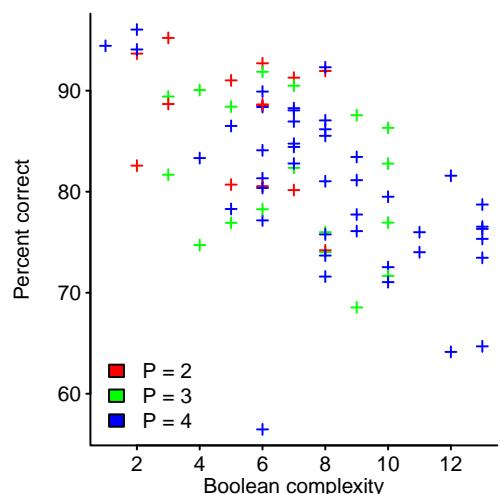


Figure 2.25: Percentage of correct answers given by one subject, against Boolean-complexity of category, colored by number of positive cases needed to define the category. Data kindly provided by Feldman.<sup>420</sup> code

#### 2.4.2.1 Categorization consistency

Receiving benefits from category usage requires some degree of usage consistency. In the case of an individual's personal categories, a person has to be able to assign the appropriate items to the correct category; within groups there has to be some level of agreement between people using the same category.

Cross-language research has found that there are very few concepts that might be claimed to be universal (they mostly relate to the human condition).<sup>1414, 1450</sup>

Human languages encode views of the world. The extent to which the language used by a person influences their thought processes continues to be hotly debated.<sup>487</sup> The proposal that language does influence thought is known as the *Sapir-Whorf* or *Whorfian* hypothesis. Some people hold what is known as the *strong language-based* view, believing that the language used does influence its speakers' conceptualization process, while people holding the so-called *weak language-based* view believe that linguistic influences occur in some cases (e.g., English contains count nouns and speakers have to pay attention to whether one or more than one item is being referred to; languages without count nouns do not require speakers to pay attention to this detail, but perhaps require them to pay attention to the shape of the object being discussed, e.g., Japanese). The *language-as-strategy* view holds that language affects speakers performance by constraining what can be said succinctly with the set of available words (a speed/accuracy trade-off, approximating what needs to be communicated in a brief sentence rather than using a longer sentence to be more accurate).

While different languages may contain different ways of describing the world, common usage patterns can be found. A study by Berlin and Kay<sup>131</sup> isolated what they called the *basic color terms* of 98 languages. They found that the number and kind of basic color terms in languages followed a consistent pattern (see Figure 2.26); while the boundaries between color terms varied, the visual appearance of the basic color terms was very similar across languages. Simulations of the evolution of color terms<sup>95</sup> suggest that it takes time for the users of a language to reach consensus on the naming of colors and over time languages accumulate more color terms.

Context plays an important role in the creation and use of categories.

A study by Bailenson, Shum, Atran, Medin and Coley<sup>83</sup> compared the categories created for two sets (US and Maya) of 104 bird species by three groups, asked US bird experts (average of 22.4 years bird watching), US undergraduates, and ordinary Itzaj (Maya Amerindians people from Guatemala). The categorization choices made by the three groups of subjects were found to be internally consistent within each group. The US experts correlated highly with the scientific taxonomy for both sets of birds, the Itzaj only correlated highly for Maya birds, and the nonexperts had a low correlation for either set of birds. The reasons given for the Maya choices varied between the expert groups; US experts were based on a scientific taxonomy, Itzaj were based on ecological justifications (the bird's relationship with its environment). Cultural differences were found in that, for instance, US subjects were more likely to generalise from songbirds, while the Itzaj were more likely to generalize from perceptually striking birds.

Context can play an important role in classification. A study by Labov<sup>799</sup> showed subjects pictures of items that could be classified as either cups or bowls (see upper plot in Figure 2.27). These items were presented in one of two contexts—a neutral context in which the pictures were simply presented and a food context (they were asked to think of the items as being filled with mashed potatoes).

Figure 2.27, lower plot, shows that as the width of the item seen was increased, an increasing number of subjects classified it as a bowl. By introducing a food context subjects responses shifted towards classifying the item as a bowl at narrower widths.

The same situation can often be viewed from a variety of different points of view (the term *frame* is sometimes used); for instance, commercial events include buying, selling, paying, charging, pricing, costing, spending, and so on. Figure 2.28 shows four ways (i.e., buying, selling, paying, and charging) of looking at the same commercial event.

A study by Jones<sup>703</sup> investigated the extent to which different developers make similar decisions when creating data structures to represent the same information... organization of struct members, what information closely associated in the data structs... Figure 12.3

### 2.4.3 Expertise

The term *expert* might be applied to a person because of their professional standing<sup>iii</sup> or because of what they can do (i.e., they know a great deal about a particular subject, or can

<sup>iii</sup> Industrial countries use professionalism as a way of institutionalising expertise.

Figure 2.26: The Berlin and Kay<sup>131</sup> language color hierarchy. The presence of any color term in a language implies the existence, in that language, of all terms below it. Papuan Dani has two terms (black and white), while Russian has eleven (Russian may also be an exception in that it has two terms for blue.) code

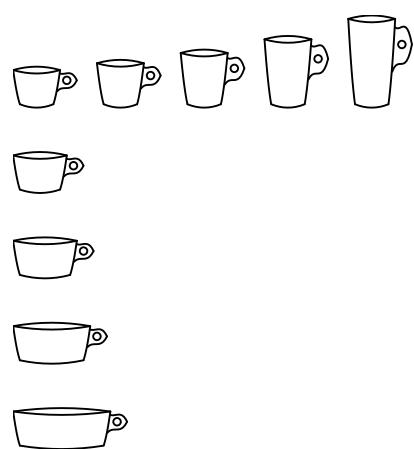
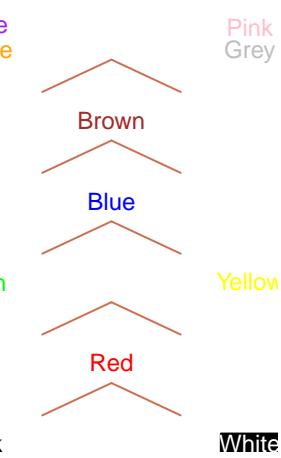


Figure 2.27: Cup- and bowl-like objects of various widths (ratios 1.2, 1.5, 1.9, and 2.5) and heights (ratios 1.2, 1.5,

perform at a qualitatively higher level than a novice in a specific domain, or some combination of the two).<sup>504</sup>

There are domains where professional experts do not perform significantly better than non-experts. For instance, in typical cases the performance of medical experts was not much greater than those of doctors after their first year of residency, although much larger differences were seen for difficult cases,<sup>7</sup> and in some cases expertise can constrain the search for solutions.<sup>1452</sup>

This section discusses expertise as a high-performance skill; something that requires many years of training and substantial numbers of individuals fail to develop proficiency.

In many fields expertise is acquired by memorizing a huge amount of domain-specific information, organizing it for rapid retrieval based on patterns that occur when problem solving within the domain and refining the problem solving process.<sup>394</sup>

Chess players were the subject of the first major study of expertise by de Groot,<sup>320</sup> and techniques used to study Chess, along with the results obtained, continue to dominate the study of expertise. In a classic study, de Groot briefly showed subjects the position of an unknown game and asked them to reconstruct it. The accuracy and speed of experts (e.g., Grand Masters) was significantly greater than non-experts when the pieces appeared on the board in positions corresponding to a game, but was not much greater when the pieces were placed at random. An explanation of the significant performance difference is that experts are faster to recognise relationships between the positions of pieces and make use of their large knowledge of positional patterns to reduce the amount of working memory needed to remember what they were briefly shown.

A study by McKeithen, Reitman, Ruster and Hirtle<sup>920</sup> gave subjects two minutes to study the source code of a program and then gave them three minutes to recall the 31 lines of the program. They were then given another two minutes to study the same source code and asked to recall the code; this process was repeated for a total of five trials.

Figure 2.29 shows the number of lines recalled by experts (over 2,000 hours of general programming experience), intermediates (just completed a programming course) and beginners (about to start a programming course) over the five trials. The upper plot are the results for the 31 line program and the lower plot a scrambled version of the program.

There is a belief that experts have some innate ability or capacity that enables them to do what they do so well. Research over the last two decades has shown that while innate ability can be a factor in performance (there do appear to be genetic factors associated with some athletic performances), the main factor in developing expert performance is time spent in *deliberate practice*;<sup>392</sup> deliberate practice does not explain everything<sup>576</sup> ...

Deliberate practice is different from simply performing the task. It requires that people monitor their practice with full concentration and receive feedback<sup>631</sup> on what they are doing (often from a professional teacher). It may also involve studying components of the skill in isolation, attempting to improve on particular aspects. The goal of this practice being to improve performance, not to produce a finished product.

Studies of the backgrounds of recognized experts, in many fields, found that the elapsed time between them starting out and carrying out their best work was at least 10 years, often with several hours of deliberate practice every day of the year. For instance, a study of violinists<sup>393</sup> (a perceptual-motor task), found that by age 20 those at the top-level had practiced for 10,000 hours, those at the next level down 7,500 hours, and those at the lowest level of expertise had practiced for 5,000 hours; similar quantities of practice were found in those attaining expert performance levels in purely mental activities (e.g., chess).

People often learn a skill for some purpose (e.g., chess as a social activity, programming to get a job) without the aim of achieving expert performance. Once a certain level of proficiency is achieved, they stop trying to learn and concentrate on using what they have learned (in work, and sport, a distinction is made between training for and performing the activity). During everyday work, the goal is to produce a product or to provide a service. In these situations people need to use well-established methods, not try new (potentially dead-end, or leading to failure) ideas to be certain of success. Time spent on this kind of practice does not lead to any significant improvement in expertise, although people may become very fluent in performing their particular subset of skills.

Expertise within one domain does not confer any additional skills within another domain,<sup>?</sup> e.g., statistics (unless the problem explicitly involves statistical thinking within the applicable

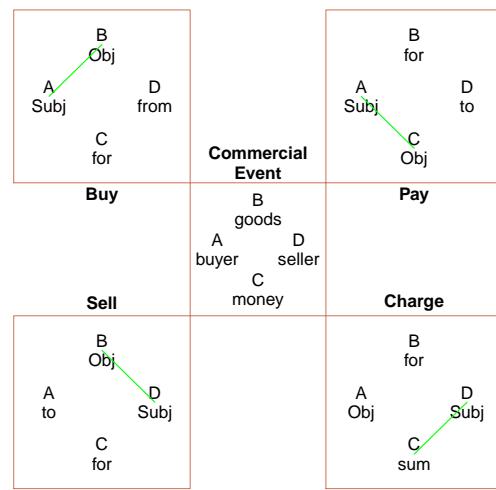


Figure 2.28: A commercial event involving a buyer, seller, money, and goods; as seen from the buy, sell, pay, or charge perspective. Based on Fillmore's code

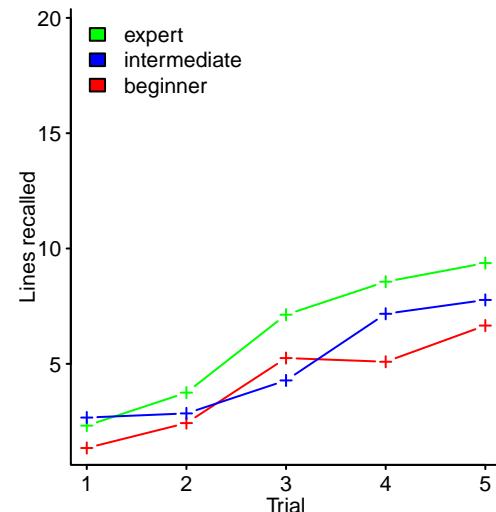
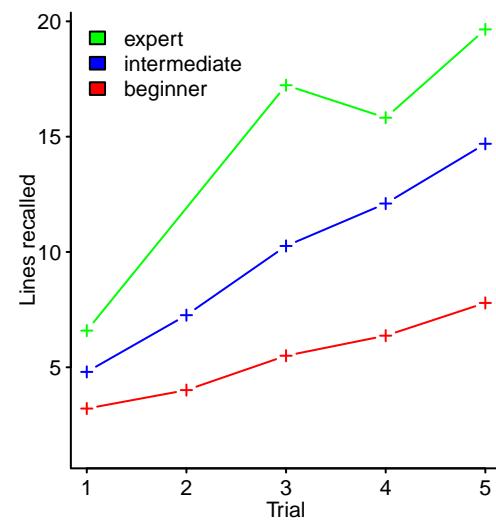


Figure 2.29: Lines of code correctly recalled after a given number of 2 minute memorization sessions; upper plot actual program, lower plot line order scrambled. Data extracted from McKeithen et al.<sup>920</sup> code

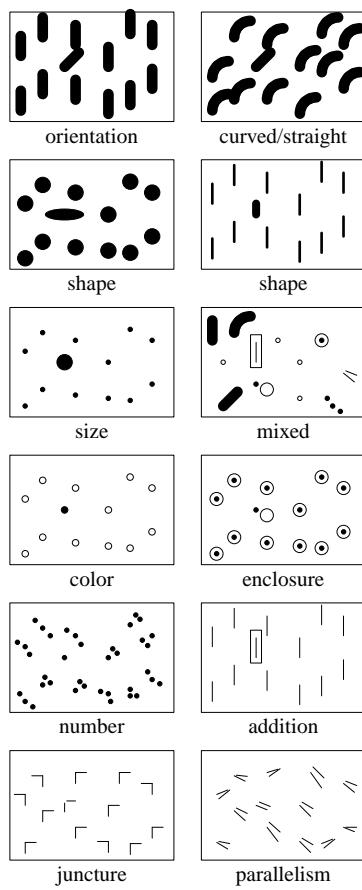


Figure 2.30: Examples of features that may be preattentively processed (parallel lines and the junction of two lines are the odd ones out). Based on Ware.<sup>1428</sup>

domain) and logic<sup>245</sup> and subjects who learned to remember long sequences of digits (after 50–100 hours of practice they could commit to memory, and recall later, sequences containing more than 20 digits) did not transfer their expertise to learning sequences of other items.<sup>240</sup>

What of individual aptitudes? In the cases studied the effects of aptitude, if there are any, have been found to be completely overshadowed by differences in experience and deliberate practice times. Willingness to spend many hours, every day, studying to achieve expert performance is certainly a necessary requirement. Does an initial aptitude or interest in a subject lead to praise from others (the path to musical and chess expert performance often starts in childhood), which creates the atmosphere for learning, or are other issues involved? IQ scores do correlate to performance during and immediately after training, but the correlation reduces over the years. The IQ scores of experts has been found to be higher than the average population, at about the level of college students.

Education can be thought of as trying to do two things (of interest to us here)—teach students skills (procedural knowledge) and providing them with information, considered important in the relevant field, to memorize (declarative knowledge).

Does attending courses in particular subjects have any measurable effect on students' capabilities, other than being able to answer questions in an exam? That is, having developed some skill in using a particular system of reasoning, do students apply it outside of the domain in which they learnt it?

A study by Lehman, Lempert, and Nisbett<sup>832</sup> measured changes in students' statistical, methodological, and conditional reasoning abilities (about everyday-life events) between their first and third years. They found that both psychology and medical training produced large effects on statistical and methodological reasoning, while psychology, medical, and law training produced effects on the ability to perform conditional reasoning; training in chemistry had no effect on the types of reasoning studied. An examination of the skills taught to students studying in these fields showed that they correlated with improvements in the specific types of reasoning abilities measured.

## 2.5 Visual processing

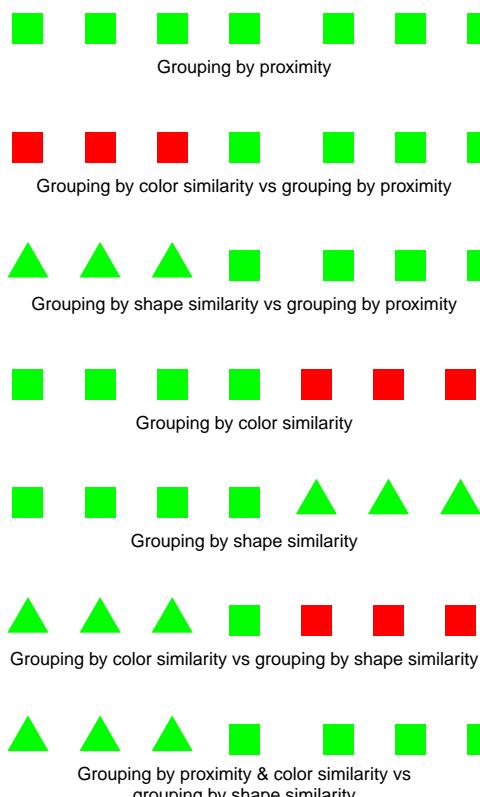


Figure 2.31: Perceived grouping of items on a line may be by shape, color or proximity. Based on kubovy et al.<sup>788</sup> code

An understanding of human visual processing is of interest to software development because it consumes cognitive resources and is a source of information input error. The 2-D image that falls on the retina does not contain enough information to build the 3-D model we see, the mind creates this model by making assumptions about how objects in our environment move and interact.<sup>626</sup>

The perceptual systems of organisms have evolved to detect information in the environment that is relevant to survival, and ignore the rest. The relevant information is about opportunities afforded by the world.

The human visual system contains various hardware systems dedicated to processing visual information; low level details are preprocessed to build structures having a high level of abstraction, e.g., lines. Preattentive processing, so called because it occurs before conscious attention,<sup>1131</sup> is automatic and apparently effortless. This preprocessing causes inputs to the visual system to appear to pop-out from their surroundings. Figure 2.30 contains examples of features that pop-out at the reader.

Preattentive processing is independent of the number of distractors; a search for the feature takes the same amount of time whether it occurs with one, five, ten, or more other distractors. However, it is only effective when the features being searched for are relatively rare. When a display contains many, distinct features (the mixed category in Figure 2.30), the pop-out effect does not occur.

The *Gestalt laws of perception* (“gestalt” means “pattern” in German, also known as the *laws of perceptual organization*)<sup>1417</sup> are based on the underlying idea that the whole is different from the sum of its parts. These so-called laws do not have the rigour expected of a scientific law, and ought to be called by some other term (e.g., principle). The Gestalt principles are preprogrammed (i.e., there is no conscious cognitive cost). The following are some commonly occurring principles

- Continuity, also known as *good continuation*: Lines and edges that can be seen as smooth and continuous are perceptually grouped together, see upper plot in Figure 2.32,

- Closure: elements that form a closed figure are perceptually grouped together, see upper plot in Figure 2.32,
- Symmetry: treating two, mirror image lines as though they form the outline of an object, see second down plot in Figure 2.32. This effect can also occur for parallel lines.
- Proximity: elements that are close together are perceptually grouped together, see second from bottom plot in Figure 2.32,
- Similarity: elements that share a common attribute can be perceptually grouped together, see lower plot Figure 2.32.
- Other: principles include grouping by connectedness, grouping by common region, and synchrony.<sup>1051</sup>

The organization of visual grouping of elements in a display, using these principles, is a common human trait. However, different people can make different choices in the perceptually grouping of the same collection of items. Figure 2.31 shows items on a horizontal line, which readers may group by shape, color or relative proximity. A study by Kubovy and van den Berg<sup>788</sup> created a model that calculated the probability of a particular perceptual grouping (i.e., shape, color, or proximity in two dimensions) being selected for a given set of items.

When mapping the prose specification of a mathematical relationship to a formula, the error rate has been found to be affected by the visual proximity of the applicable words.<sup>809</sup>

A number of studies have found that people are more likely to notice the presence of a distinguishing feature than the absence of a distinguishing feature. This characteristic affects performance when searching for an item when it occurs among visually similar items. It can also affect reading performance—for instance, substituting an e for a c is more likely to be noticed than substituting a c for an e.

A study by Treisman and Souther<sup>1361</sup> found that visual searches were performed in parallel when the target included a unique feature (i.e., search time was not affected by the number of background items), but were performed serially when the target had a unique feature missing (i.e., search time was proportional to the number of background items).

What is a unique feature? Subjects searched for circles that differed in the presence or absence of a gap (see Figure 2.33). The results showed that subjects were able to locate a circle containing a gap, in the presence of complete circles, in parallel. However, searching for a complete circle, in the presence of circles with gaps, was performed serially. In this case the gap was the unique feature; performance also depended on the proportion of the circle taken up by the gap...

The visual aspect of creating a software system involves the reading code and text. Research on the cognitive processes involved in reading prose written in human languages has uncovered the basic processes and various models have been built.<sup>1146</sup>

During reading, a person's eyes make short rapid movements, known as *saccades*, taking 20 ms to 50 ms to complete; a saccade typically moves the eyes forward 6 to 9 characters. No visual information is extracted during a saccade and readers are not consciously aware of them. Between saccades the eyes are stationary, typically for 200 ms to 250 ms, these stationary periods are known as *fixations*; a study of consumer eye movements<sup>1090</sup> while comparing multiple brands found a fixation duration of 354 ms when subjects were under high time pressure and 431 ms when under low time pressure.

Individual readers can exhibit considerable variations in performance, a saccade might move the eyes by one character, or 15 to 20 characters; fixations can be shorter than 100 ms or longer than 400 ms (there is also variation between languages<sup>1038</sup>). The content of the fixated text has a strong effect on performance.

The eyes do not always move forward during reading—10% to 15% of saccades move the eyes back to previous parts of the text. These backward movements, called *regressions*, are caused by problems with linguistic processing (e.g., incorrect syntactic analysis of a sentence) and oculomotor error (for instance, the eyes overshooting their intended target).

Saccades are necessary because the eyes' field of view is limited. Light entering an eye hits light-sensitive cells in the retina, where cells are not uniformly distributed. The visual field (on the retina) can be divided into three regions: foveal (the central 2°, measured from the

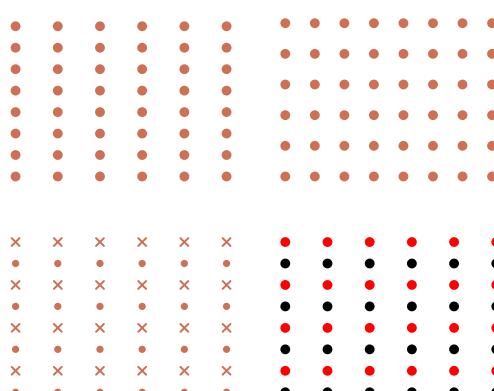
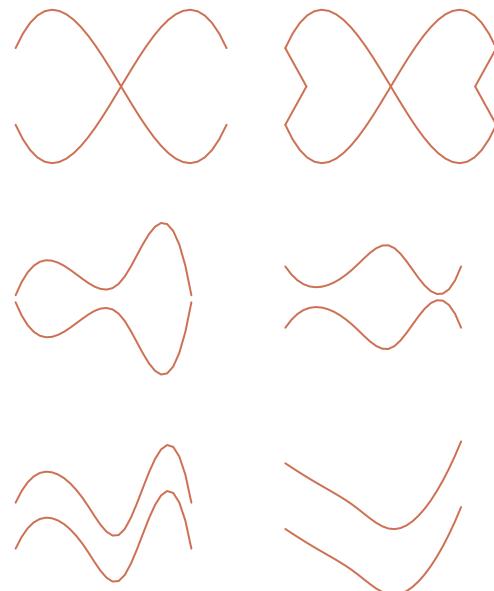
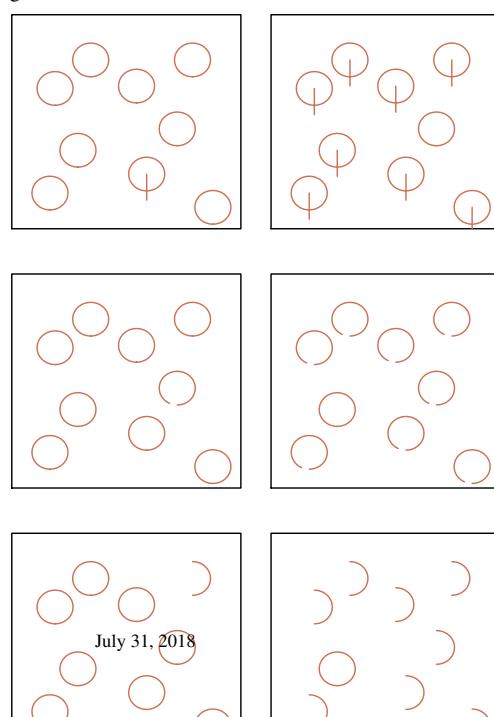


Figure 2.32: Continuity—upper left plot is perceived as two curved lines; Closure—when the two perceived lines are joined at their end (upper right), the perception changes to one of two cone-shaped objects; Symmetry and parallelism—where the direction taken by one line follows the same pattern of behavior as another line; Proximity—the horizontal distance between the dots in the lower left plot is less than the vertical distance, causing them to be perceptually grouped into lines (the relative distances are reversed in the right plot); Similarity—a variety of dimensions along which visual items can differ sufficiently to cause them to be perceived as being distinct; rotating two line segments by 180° does not create as big a perceived difference as rotating them by 45°; TODO look good. [code](#)



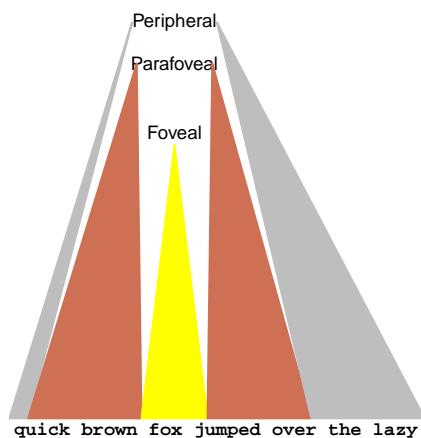


Figure 2.34: The foveal, parafoveal and peripheral vision regions when three characters visually subtend 3°. Based on Schotter et al.<sup>1213</sup> code

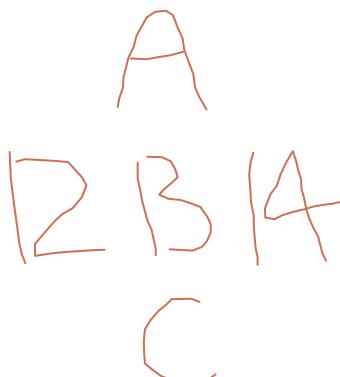
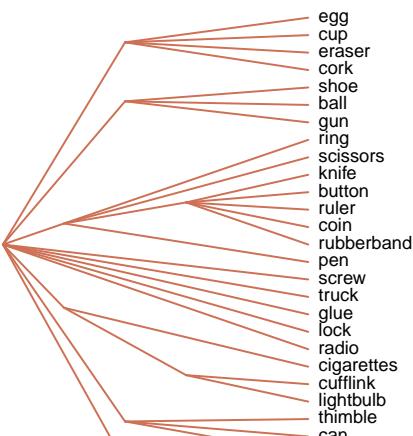


Figure 2.35: Local context can change the interpretation given to the surrounding symbols. code

thimble	cigarettes	radio
	lightbulb	
can		glue
flashlight	cufflink	lock
	book	
cork	envelope	truck
eraser		
cup	rubberband	screw
	ruler	
egg	coin	pen
button		ring
gun	knife	scissors
ball		shoe



front of the eye looking toward the retina), parafoveal (extending out to 5°), and peripheral (everything else). Letters become increasingly difficult to identify as their angular distance from the center of the fovea increases.

Two processes during the fixation period: identifying the word (or sequence of letters forming a partial word) and planning the next saccade, when to make it and where to move the eyes. Reading performance is speed limited by the need to plan and perform saccades (removing the need to saccade by presenting words at the same place on a display, there is a threefold speed increase in reading aloud and a two-fold speed increase in silent reading). The time needed to plan and perform a saccade is approximately 180 ms to 200 ms (known as the *saccade latency*), which means that the decision to make a saccade occurs within the first 100 ms of a fixation.

The contents of the parafoveal region are partially processed during reading and this increases a reader's perceptual span. When reading words written using alphabetic characters, the perceptual span extends from 3 to 4 characters on the left of fixation to 14 to 15 letters to the right of fixation. This asymmetry in the perceptual span is a result of the direction of reading, attending to letters likely to occur next being of greater value. Readers of Hebrew (which is read right-to-left) have a perceptual span that has opposite asymmetry (in bilingual Hebrew/English readers the direction of the asymmetry depends on the language being read, showing the importance of attention during reading).<sup>1145</sup>

Characteristics used by the writing system affect the asymmetry of the perceptual span and its width, e.g., the span can be smaller for Hebrew than English (Hebrew words can be written without the vowels, requiring greater effort to decode and plan the next saccade). It is also much smaller for writing systems that use ideographs, such as Japanese (approximately 6 characters to the right) and Chinese.

The perceptual span is not hardwired, but is attention-based. The span can become smaller when the fixated words are difficult to process. Also, readers extract more information in the direction of reading when the upcoming word is highly predictable (based on the preceding text).

Models of reading have achieved some level of success include: Mr. Chips,<sup>831</sup> an ideal-observer model of reading (it is not intended to model of how humans read, but to establish the pattern of performance when optimal use is made of available information) which attempts to calculate the distance, in characters, of the next saccade; it combines information from visual data obtained by sampling the text through a retina, lexical knowledge of words and their relative frequencies, and motor knowledge of the statistical accuracy of saccades and uses the optimization principle of entropy minimization. SWIFT:<sup>387</sup> attempts to provide a realistic model of saccade generation, E-Z Reader:<sup>1101</sup> attempts to account for how cognitive and lexical processes influence the eye movements of skilled readers; it can handle the complexities of *garden path* sentences.<sup>1151iv</sup>

English text is written on lines down a page and read left to right. The order in which the components of a formula are read depends on its contents, with the visual processing of subexpressions by experienced users driven by the mathematical syntax,<sup>682,683</sup> and the extraction of syntax happening in parallel.<sup>1207</sup>

A study by Adelson<sup>7</sup> investigated the organization present in Subject's recall order of previously remembered lines of code. Subjects saw a total of 16 lines of code, one line at a time, and were asked to memorise each line for later recall. The lines were taken from three simple programs (two containing five lines each and one containing six lines), the order being randomised for each subject. Five subjects were students who had recently taken a course using the language used and five subjects had recently taught a course using the language.

Subjects were free to recall the previously seen lines in any order. An analysis of the recall order showed that students grouped lines by syntactic construct (e.g., loop, function header), while the course teachers recalled the lines in the order in which they appeared in the three programs (i.e., they reconstructed three programs from the 16 randomly ordered lines); see Figure 2.37.

A study by Pelli, Burns, Farrell, and Moore<sup>1072</sup> found that 2,000 to 4,000 trials were all that was needed for novice readers to reach the same level of efficiency as fluent readers in the

<sup>iv</sup> In "Since Jay always jogs a mile seems like a short distance." readers experience a disruption that is unrelated to the form or meaning of the individual words; the reader has been led down the syntactic garden path of initially parsing the sentence such that *a mile* is the object of *jogs* before realizing that *a mile* is the subject of *seems*.

letter-detection task (efficiency was measured by comparing human performance compared to an ideal observer). They tested subjects aged 3 to 68 with a range of different (and invented) alphabets (including Hebrew, Devanagari, Arabic, and English). Even fifty years of reading experience, over a billion letters, did not improve the efficiency of letter detection. They also found this measure of efficiency was inversely proportional to letter perimetric complexity (defined as, inside and outside perimeter squared, divided by *ink* area).

Studies have found that peoples memory for objects within their visual field of view is organized according to the relative positions of the objects. For instance, a study by McNamara, Hardy, and Hirtle<sup>923</sup> gave subjects two minutes to memorize the location of objects on the floor of a room (see upper plot in Figure 2.36). The objects were then placed in a box and subjects were asked to place the objects in their original position. The memorize/recall cycle was repeated, using the same layout, until the subject could place all objects in their correct position.

The order in which each subject recalled the location of objects was used to create a hierarchical tree (one for each subject). The resulting trees (see lower plot in Figure 2.36) showed how subjects' spatial memory of the objects seen had a hierarchical organization, with the spatial distance between items being a significant factor in its structure.

Choice of display font is something that many developers are completely oblivious to. The use of Roman, rather than Helvetica (or serif vs. sans serif), is often claimed to increase reading speed and comprehension. The issues involved in selecting fonts are covered in a report detailing "Font Requirements for Next Generation Air Traffic Management Systems".<sup>183</sup>

Vision provides information about what people are thinking about; our gaze follows shifts of visual attention. Tracking a subject's eye movements and fixations when viewing images or real-life scenes is an established technique in fields such as marketing and reading research; this technique is now starting to be used to research developer code reading behavior (see Figure 2.38).

## 2.6 Reasoning

The use case for reasoning is extracting information from available data; adding constraints on the data (e.g., known behaviors) can increase the quantity of information that can be extracted.

What survival advantages does an ability to reason provide, or is it primarily an activity WEIRD people need to learn to pass school tests?

The kinds of questions asked in studies of reasoning appear to be uncontentious. However, studies<sup>871, 1219</sup> of reasoning using illiterate subjects from remote parts of the world received answers to verbal reasoning problems that were based on personal experience and social norms, rather than the western ideal of logic. The answers given by subjects, in the same location, who had received several years of schooling were much more likely to match those demanded by mathematical logic; the subjects had learned how to be WEIRD and *play the game*. The difficulty experienced by those learning formal logic suggests that there is no innate capacity for this task (innate capacity enables the corresponding skill to be learned quickly and easily). The human mind is a story processor, not a logic processor.<sup>571</sup>

The Wason selection task<sup>1431</sup> is to studies of reasoning like the fruit fly is to studies of genetics. Wason's study was first published in 1968 and considered mathematical logic to be the norm against which human reasoning performance should be judged. The reader might like to try this selection task:

- Figure 2.39 depicts a set of four cards, of which you can see only the exposed face but not the hidden back. On each card, there is a number on one of its sides and a letter on the other.
- Below there is a rule which applies only to the four cards. Your task is to decide which, if any, of these four cards you must turn in order to decide if the rule is true.
- Don't turn unnecessary cards. Tick the cards you want to turn.

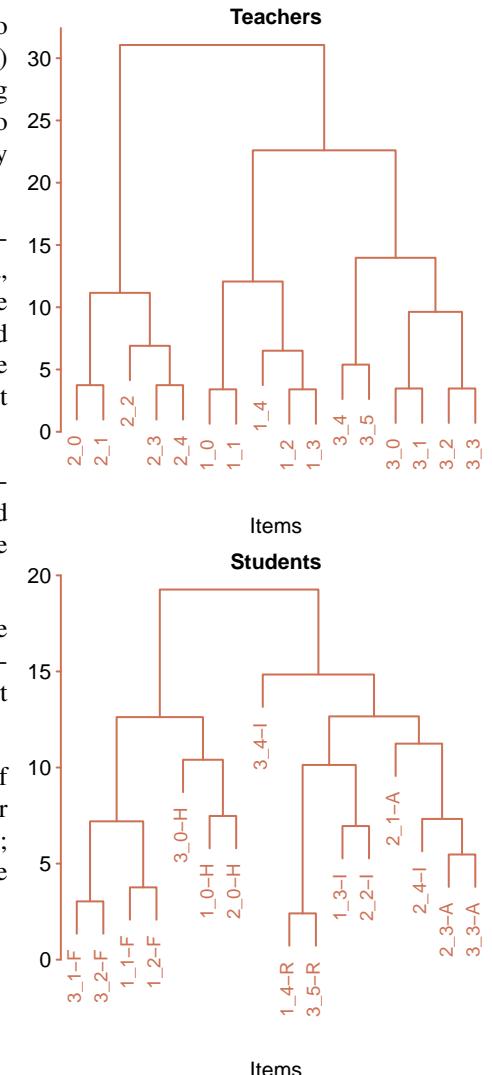


Figure 2.37: Hierarchical clustering of statement recall order, averaged over teachers and then students; label names are: *programlist* – *statementkind*, *wherestatementkindmightbeafunctionhead*

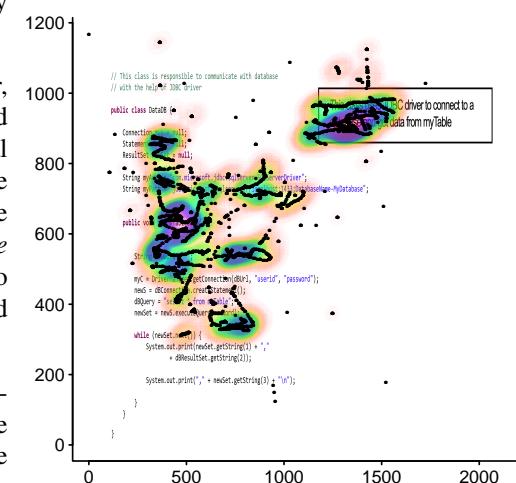


Figure 2.38: Heat map of one subject's cumulative fixations (black dots) on a screen image. Data kindly provided by Ali.<sup>20</sup> code

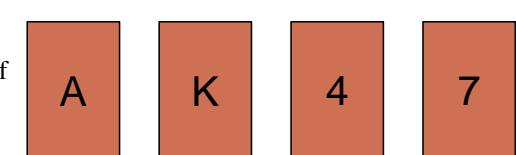


Figure 2.39: The four cards used in the Wason selection task. Based on Wason.<sup>1431</sup> code July 3, 2009

**Rule:** If there is a vowel on one side, then there is an even number on the other side.

**Answer: ?** The failure of many subjects to give the expected answer (i.e., the one derived using mathematical logic<sup>v</sup>) surprised many researchers and over the years a wide variety of explanations, experiments, thesis and books have attempted to explain what is going on. Explanations for subject behavior include: human reasoning is tuned to detecting people who cheat<sup>292</sup> within a group where mutual altruism is the norm, interpreting the wording of questions pragmatically based on how natural language is used rather than as logical formula<sup>619</sup> (i.e., assuming a social context; people are pragmatic virtuosos rather than logical defectives), and adjusting norms to take into account cognitive resource limitations (i.e., computational and working memory) or a rational analysis approach.<sup>1026</sup>

Dual process theories<sup>1268, 1298, 1299</sup> treat people as having two systems: unconscious reasoning and conscious reasoning.

Understanding spoken language requires reasoning about what is being discussed<sup>934</sup> and in a friendly shared environment it is possible to fill in the gaps by assuming that what is said is relevant,<sup>1286</sup> with no intent to trick. In an adversarial context sceptical reasoning of the mathematical logic kind is useful for enumerating all possible interpretations of what has been said.

Outside of classroom problems, a real world context in which people explicitly reason is decision-making and here “fast and frugal algorithms”<sup>494, 497</sup> which provide answers quickly within the often, limited, constraints of time and energy. Context and semantics are crucial inputs to the reasoning process.<sup>1307</sup>

Reasoning and decision-making appear to be closely related. However, reasoning researchers tied themselves to using the norm of mathematical logic for many decades and created something of research ghetto,<sup>1309</sup> while decision-making researchers have been involved in explaining real-world problems.

Some Wason task related studies used a dialogue protocol (i.e., subjects’ discuss their thoughts about the problem with the experimenter) and transcriptions<sup>1308</sup> of these studies read like people new to programming having trouble understanding what it is they have to do to solve a problem by writing code.

People have different aptitudes and this can result in them using different strategies to solve the same problem,<sup>1298</sup> e.g., an interaction between a subject’s verbal and spatial ability and the strategy used to solve linear reasoning problems.<sup>1311</sup> However, a person having high spatial ability, for instance, does not necessarily use a spatial strategy. A study by Roberts, Gilmore, and Wood<sup>1164</sup> asked subjects to solve what appeared to be a spatial problem (requiring the use of a very inefficient spatial strategy to solve). Subjects with high spatial ability used non-spatial strategies, while those with low spatial ability used a spatial strategy. The conclusion made was that those with high spatial ability were able to see the inefficiency of the spatial strategy and selected an alternative strategy, while those with less spatial ability were unable to make this evaluation.

A study by Bell and Johnson-Laird<sup>121</sup> investigated the effect of the kind of questions asked on reasoning performance. Subjects had to give yes/no responses to two kinds of questions, asking about what is possible or what is necessary. The hypothesis was that subjects would find it easier to infer a *yes* answer to a question about what is possible, compared to one about what is necessary, because only one instance needs to be found (all instances need to be checked to answer *yes* to a question about necessity). For instance, in a game in which only two can play and the following information:

If Allan is in then Betsy is in.  
If Carla is in then David is out.

answering *yes* to the question “Can Betsy be in the game?” (a possibility) is easier than giving the same answer to “Must Betsy be in the game?” (a necessity); see Table 2.4.

However, subjects would find it easier to infer a *no* answer to a question about what is necessary, compared to one about what is possible, because only one instance needs to be found, whereas all instances need to be checked to answer *no* to a question about possibility. For instance, in another two person game and the following information:

---

<sup>v</sup> blah...

Question	Correct yes	Correct no
is possible	91%	65%
is necessary	71%	81%

Table 2.4: Percentage of correct answers to the two kinds of questions and two kinds of response. Data from Bell et al.<sup>121</sup>

If Allan is out then Betsy is out.

If Carla is out then David is in.

answering *no* to the question “Must Betsy be in the game?” (a necessity) is easier than giving the same answer to “Can Betsy be in the game?” (a possibility).

**Conditionals in English** In natural languages the conditional clause generally precedes the conclusion, in a conditional statement;<sup>545</sup> an example where the conditional follows the conclusion is “I will leave, if you pay me” given as the answer to the question “Under what circumstances will you leave?”. In one study of English<sup>454</sup> the conditional preceded the conclusion in 77% of written material and 82% of spoken material. There is a lot of variation in the form of the conditional.<sup>137,219</sup>

**Alfred North Whitehead** It is a profoundly erroneous truism . . . that we should cultivate the habit of thinking of what we are doing. The precise opposite is the case. Civilization advances by extending the number of important operations which we can perform without thinking about them.

## 2.6.1 Deductive reasoning

The following are some factors that have been found to affect peoples performance in solving deductive reasoning problems:

- Belief bias: people are more willing to accept a conclusion, derived from given premises, that they believe to be true than one they believe to be false; see Table 2.5. A study by Evans, Barston, and Pollard<sup>400</sup> gave subjects two premises and a conclusion and asked them to state whether the conclusion was true or false (based on the premises given; the conclusions were rated as either believable or unbelievable by a separate group of subjects).

Status-context	Example	Accepted
Valid-believable	No Police dogs are vicious Some highly trained dogs are vicious Therefore, some highly trained dogs are not police dogs	88%
Valid-unbelievable	No nutritional things are inexpensive Some vitamin tablets are inexpensive Therefore, some vitamin tablets are not nutritional things	56%
Invalid-believable	No addictive things are inexpensive Some cigarettes are inexpensive Therefore, some addictive things are not cigarettes	72%
Invalid-unbelievable	No millionaires are hard workers Some rich people are hard workers Therefore, some millionaires are not rich people	13%

Table 2.5: Percentage of subjects accepting that the stated conclusion could be logically deduced from the given premises. Based on Evans et al.<sup>400</sup>

- Form of premise: a study by Dickstein<sup>351</sup> measured subjects performance on the 64 possible two premise syllogisms (a premise being one of the propositions: *All S are P*, *No S are P*, *Some S are P*, and *Some S are not P*). For instance, the following syllogisms show the four possible permutations of three terms (the use of S and P is interchangeable):

$$\begin{array}{llll} \text{All M are S} & \text{All S are M} & \text{All M are S} & \text{All S are M} \\ \text{No P are M} & \text{No P are M} & \text{No M are P} & \text{No M are P} \end{array}$$

The results showed that performance was affected by the order in which the terms occurred in the two premises of the syllogism.

The order in which the premises are processed may affect the amount of working memory needed to reason about the syllogism, which in turn can affect human performance.<sup>500</sup>

- Individual differences<sup>78</sup> ...

## 2.6.2 Linear reasoning

The ability to make relational comparisons provides a number of benefits, including selecting which of two areas contains the largest amount of food. Some animals, including humans, have a biologically determined representation of numbers, including elementary arithmetic operations, what one researcher has called the *number sense*.<sup>334</sup>

Being able to make relational decisions is a useful skill for animals living in hierarchical social groups where aggression is used to decide status.<sup>1066</sup> Aggression is best avoided, as it can lead to injury; the ability to make use of relative dominance information (obtained by watching interactions between other members of the group) may remove the need for aggressive behavior during an encounter between two group members who have not recently contested dominance (i.e., there is nothing to be gained in aggression towards a group member who has previously been seen to dominate a member who is dominant to yourself).

The use of relational operators have an obvious interpretation in terms of linear syllogisms. A study by De Soto, London, and Handel<sup>324</sup> investigated a task they called *social reasoning*, using the relations *better* and *worse*. Subjects were shown two relationship statements involving three people, and a possible conclusion (e.g., “Is Mantle worse than Moskowitz?”) and had 10 seconds to answer “yes”, “no”, or “don’t know”. The British National Corpus<sup>828</sup> lists *better* as appearing 143 times per million words, while *worse* appears under 10 times per million words and is not listed in the top 124,000 most used words.

	<b>Relationships</b>	<b>Correct %</b>		<b>Relationships</b>	<b>Correct %</b>
1	A is better than B B is better than C	60.5	5	A is better than B C is worse than B	61.8
2	B is better than C A is better than B	52.8	6	C is worse than B A is better than B	57.0
3	B is worse than A C is worse than B	50.0	7	B is worse than A B is better than C	41.5
4	C is worse than B B is worse than A	42.5	8	B is better than C B is worse than A	38.3

Table 2.6: Eight sets of premises describing the same relative ordering between A, B, and C (peoples names were used in the study) in different ways, followed by the percentage of subjects giving the correct answer. Based on De Soto et al.<sup>324</sup>

There are two patterns of performance behavior visible in Table 2.6.

- A higher percentage of correct answers were given when the direction was better-to-worse (case 1), than mixed direction (case 2, 3), and were least correct in the direction worse-to-better (case 4),
- A higher percentage of correct answers were given when the premises stated an end term (better or worse) followed by the middle term, than a middle term followed by an end term.

A second experiment in the same study gave subjects printed statements about people. For instance, “Tom is better than Bill”. The relations used were *better*, *worse*, *has lighter hair*, and *has darker hair*. The subjects had to write the peoples names in two of four possible boxes; two arranged horizontally and two arranged vertically.

The results showed 84% of subjects selecting a vertical direction for better/worse, with better at the top (which is consistent with the *up is good* usage found in English metaphors<sup>803</sup>). In the case of lighter/darker 66% of subjects used a horizontal direction, with no significant preference for left-to-right or right-to-left.

A third experiment in the same study used the relations *to-the-left* and *to-the-right*, and *above* and *below*. The above/below results were very similar to those for better/worse. The left-right results showed that subjects learned a left-to-right ordering better than a right-to-left ordering.

The results of this study show the effect that operand order has on the accuracy of peoples responses. However, the interpretation placed on the operator also plays a significant role. Without knowing what interpretation a reader is likely to give to the operands and operators in the following two (logically equivalent) conditional expressions, for instance, it is not possible to select the one that is most likely to minimize incorrect reasoning on the part of readers.

```
if ((x <= y) && (x => z))
if ((x >= z) && (x <= y))
```

ACCU relational if-condition study...

Subject performance on linear reasoning improves the greater the distance between the items being compared; this *distance effect* is discussed in Section 2.7.

### 2.6.3 Causal reasoning

A question often asked by developers, while reading source, is “what causes this event/situation to occur?” Causal questions such as this are also common in everyday life. However, there has been relatively little mathematical research on causality (Pearl<sup>1068</sup> is the standard reference; statistics deals with correlation) and little psychological research on causal reasoning.<sup>1271</sup> It is sometimes possible to express a problem in either a causal or conditional form. A study by Sloman, and Lagnado<sup>1272</sup> gave subjects one of the following two reasoning problems and associated questions:

- Causal argument form:

```
A causes B
A causes C
B causes D
C causes D
D definitely occurred
```

with the questions: “If B had not occurred, would D still have occurred?”, or “If B had not occurred, would A have occurred?”.

- Conditional argument form:

```
If A then B
If A then C
If B then D
If C then D
D is true
```

with the questions: “If B were false, would D still be true?”, or “If B were false, would A be true?”.

Table 2.7 shows that subject performance depended on the form in which the problem was expressed (more cases to be added...).

Question	Causal	Conditional
D holds?	80%	57%
A holds?	79%	36%

Table 2.7: Percentage “yes” responses to various forms of questions (based on 238 responses). Based on Sloman et al<sup>1272</sup>

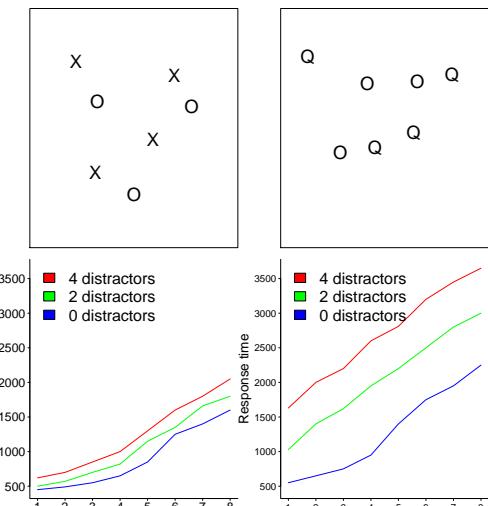


Figure 2.40: Average time (in milliseconds) taken for subjects to enumerate O's in a background of X or Q distractors. Based on Trick and Pylyshyn.<sup>1362</sup> code

interpretation of causal verbs<sup>512, 1267</sup> ...

## 2.7 Number processing

Having a sense of quantity and being able to judge the relative size of two quantities provides a number of survival benefits, including being able to perform an approximate number of

repetitions of some task (e.g., pressing a bar, see Figure 2.3) and deciding which of two clusters of food is the largest.

Being able to quickly enumerate small quantities is sufficiently useful for the brain to support for preattentive processing of up to four or so items.<sup>1362</sup> When asked to enumerate how many dots are visible in a well-defined area subjects' response time depends on the number of dots; between one and four dots performance varies between 40 ms to 100 ms per dot, but with five or more dots performance varies between 250 ms to 350 ms per dot. The faster process is known as *subitizing* (people effortlessly see the number of dots), while the slower process is called *counting*.

Studies have found that a variety of animals can make use of an approximate mental number system (sometimes known as the *number line*; see Figure 2.3); the extent to which brains have a built-in number line or existing neurons are repurposed through learning is an active area of research.<sup>332,1023</sup> Humans are the only creatures known to have a second system, one that can be used to represent numbers exactly: language.

A study by van Oeffelen and Vos<sup>1391</sup> investigated subject's ability to estimate the number of dots in a briefly (100 ms, i.e., not enough time to be able to count the dots) displayed image. Subjects were given a target number and had to answer yes/no whether they thought the image they saw contained this number of dots. Figure 2.41 shows the probability of a correct answer for various target numbers and a given difference between target number and number of dots displayed.

What are the operating characteristics of the approximate number system? The characteristics that have most occupied researchers are the scale used (e.g., linear or logarithmic), the impact of number magnitude on cognitive performance and when dealing with two numbers the effect of their relative difference in value on cognitive performance.<sup>334</sup>

Studies of how single digit numbers are mentally represented<sup>337</sup> have found a linear scale used by subjects from western societies and a logarithmic scale for subjects from indigenous cultures that have not had formal schooling.

Engineering and science sometimes deal with values spanning many orders of magnitude, a range that people are unlikely to encounter in everyday life. How do people mentally represent large value ranges?

A theoretical analysis,<sup>1089</sup> found that a logarithmic scale minimized representation error, based on the probability of a value occurring having a power-law distribution, assuming relative change is the psychologically-relevant measure, and that noise is present in the signal.

A study by Landy, Charlesworth and Ottmar<sup>810</sup> asked subjects to click the position on a line (labeled at the left end with one thousand and at the right end with one billion) which they thought was the appropriate location for each of the 182 values they saw (selected from 20 evenly spaced values between one thousand and one million, and 23 evenly spaced values between one million and one billion).

Various patterns occurred in subject responses; the top left plot in Figure 2.42 shows one of the most common. Most subjects placed one million at the halfway point (as-if using a logarithmic scale), placing values below/above a million on separate linear scales. Landy et al developed a model based on Category Adjustment theory,<sup>7</sup> where subjects selected a category boundary (e.g., one million, creating the categories: the thousands and the millions), a location for the category boundary along the line and a linear(ish)<sup>vi</sup> mapping of values to relative position within their respective category.

Studying the learning and performance of simple arithmetic operations has proven complicated;<sup>447,1393</sup> models of simple arithmetic performance<sup>823</sup> have been built. Working memory capacity has an impact on the time taken to perform mental arithmetic operations and the likely error rate, e.g., remembering carry or borrow quantities during subtraction,<sup>662</sup> the human language used to think about the problem is also has an impact.<sup>661</sup>

How do people compare multi-digit integer constants? For instance, do they compare them digit by digit (i.e., a serial comparison), or do they form two complete values before comparing their magnitudes (the so-called *holistic* model)? Studies show that the answer depends on how the comparisons are made, with results consistent with the digit by digit<sup>1485</sup> and holistic<sup>336</sup> approaches being found.

Other performance related behaviors include:

<sup>vi</sup> Category Adjustment theory supports curvaceous lines.

- *split effect*: taking longer to reject false answers that are close to the correct answer (e.g.,  $4 \cdot 7 = 29$ ) than those that are further away (e.g.,  $4 \cdot 7 = 33$ ),
- *associative confusion effect*: answering a different question from the one asked (e.g., giving 12 as the answer to  $4 \cdot 8 = ?$ , which would be true had the operation been addition),
- plausibility judgments:<sup>834</sup> using of a rule rather than retrieving a fact from memory to verify the answer to a question; for instance, adding an odd number to an even number always produces an odd result,

A study by LeFevre and Liu<sup>830</sup>...

### 2.7.1 Problem size and symbolic distance effect

The time taken to produce an answer and the error rate increase as the numeric value of operands increases (e.g., subjects are slower to solve  $9 + 7$  than  $2 + 3$ ); this is known as the *problem size effect*.

A study by Campbell<sup>213</sup> subject performance when multiplying numbers between two and nine, and dividing a number (that gave an integer result) by a number between two and nine. Figure 2.43 shows that the number of errors generally increases with operand and result value (both operands have the same value appears to be a special case, see blue points). The time taken for subjects to add<sup>1056</sup> or multiply<sup>1055</sup> two single digit values, e.g.,  $p \cdot q$ , is proportional to  $\min(p, q)$  (the time taken to confirm an answer follows a similar pattern).

The *symbolic distance effect* is a general effect that occurs when people compare two items sharing a distance-like characteristic which can be easily estimated; the larger the distance between two items the faster people are likely to respond to a comparison question involving this characteristic. For instance, comparisons of social status<sup>250</sup> and geographical distance<sup>622</sup> (also see Figure 2.36). Inconsistencies between the symbolic distance and actual distance can increase the error rate,<sup>600</sup> e.g., is the following relation true?  $3 > 5$ .

In a study by Tzelgov, Yehene, Kotler and Alon<sup>1374</sup> subjects trained one-hour per day for six days, learning the relative order of nine graphical symbols. A symbolic distance effect was seen in subject performance, when answering questions about relative graphical symbol order.

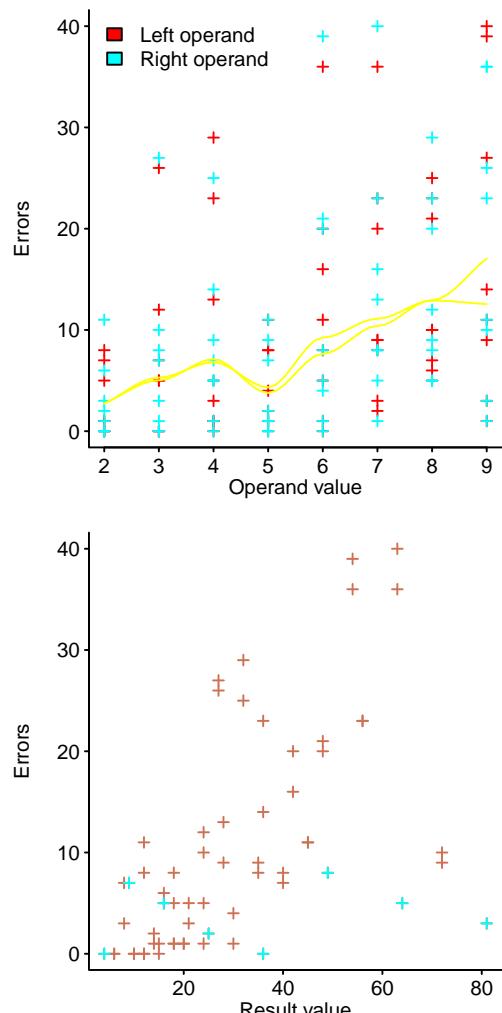


Figure 2.43: Number of errors, in 132 simple multiplication trials (e.g.,  $3 \cdot 7$ ), upper plot shows operand values (a loess fit in yellow) and lower plot result value (points where both operands have the same value are in blue). Data from Campbell.<sup>213</sup> [code](#)

## 2.8 Human factors

Not only do Humans come in various shapes and sizes, an individual's performance can significantly change from one moment to the next.

A study by Remington, Yuen and Pashler<sup>1154</sup> compared subject performance between using a GUI and a command line (with practice, there was little improvement in GUI performance, but command line performance continued to improve and eventually overtook GUI performance). Figure 2.44 shows the command line response time for one subject over successive blocks of trials and a fitted loess line.

### 2.8.1 People make mistakes

Evolution, in nature, is driven by survival of the fittest, i.e., it is based on relative performance; a consistent flawless performance is not only unnecessary, but a waste of precious resources.

Socially, making mistakes is an accepted fact of life and people are given opportunities to correct mistakes, if that is considered necessary.

Reason<sup>1148</sup> is a readable introduction to human error.

For a given task, obtaining information on the kinds of mistakes that are likely to be made (e.g., entering numeric codes on a keyboard<sup>1043</sup>) and modeling the behavior (e.g., subtraction mistakes<sup>1393</sup> made by children learning arithmetic) requires a lot of effort, even for simple tasks. Modeling the mistakes people make has proven to be very difficult. Researchers are still working to build good models<sup>1281</sup> for the apparently simple task of text entry.

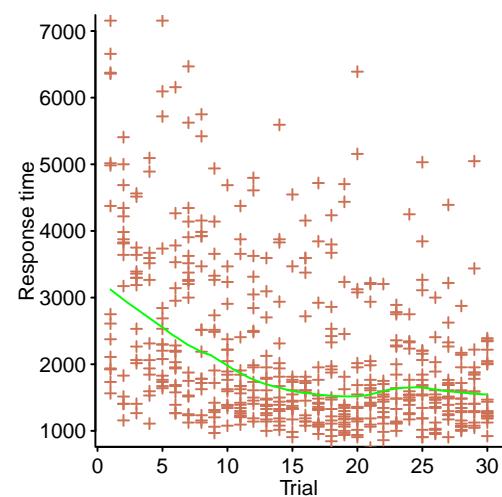


Figure 2.44: One subject's response time over successive blocks of command line trials and fitted loess (in green). Data kindly provided by Remington.<sup>1154</sup> [code](#)

One technique for increasing the number of errors made by subjects in an experiment is to introduce factors that will increase the number of mistakes made. For instance, under normal circumstances the letters/digits viewed by developers are clearly visible and the viewing time is not constrained. In experiments run under these conditions subjects make very few errors. To obtain enough data to calculate letter similarity/confusability, studies<sup>974</sup> have to show subjects images of single letters/digits that have been visually degraded in some way, or given a limited amount of time to make a decision, or both until a specified error rate is achieved.<sup>1357</sup> While such experiments may provide the only available information on the topic of interest, their ecological validity has to be addressed (compared to say asking subjects to rate pairs of letters for similarity<sup>1260</sup>).

How often do people make mistakes?

A lower bound on human error rate, when performing over an extended period, is probably that of astronauts in space; making an error during a space mission can have very serious consequences and a huge amount of resources are devoted to astronaut training. NASA maintains several databases of errors made by operators during simulation training and actual missions; measurements of human error rates, for different missions, of between  $1.9 \cdot 10^{-3}$  and  $1.05 \cdot 10^{-4}$  have occurred.<sup>228</sup>

Touch typists, who are performing purely data entry:<sup>904</sup> with no error correction 4% (per keystroke), typing nonsense words (per word) 7.5%.

A number of human reliability analysis methods<sup>229</sup> for tasks in safety critical environments are available. The Cognitive Reliability Error Analysis Model (CREAM) is widely used; Calhoun et al<sup>210</sup> work through a calculation of the probability of an error during the International Space Station ingress procedure, using CREAM.

What causes error rates to increase?... Are some people more error prone than others?...

How do people respond when their mistakes are discovered?

A study by Jørgensen and Moløkken<sup>718</sup> interviewed employees, from one company, with estimation responsibilities. Analysis of the answers showed a strong tendency for people to perceive factors outside their control as important reasons for inaccurate estimates, while factors within their control were typically cited as reasons for accurate estimates.

## 2.8.2 Cognitive effort

When attempting to solve a problem, a person's cognitive system makes cost/accuracy trade-offs. The details of how it forms an estimate of the value, cost, and risk associated with an action, and carries out the trade-off analysis is not known (various models have been proposed<sup>543</sup>). An example of the effects of these trade-offs is provided by a study by Fu and Gray,<sup>461</sup> where subjects had to copy a pattern of colored blocks (on a computer-generated display). Remembering the color of the block to be copied and its position in the target pattern created a memory effort; a perceptual-motor effort was introduced by graying out the various areas of the display where the colored blocks were visible; these grayed out areas could be made temporarily visible using various combinations of keystrokes and mouse movements. Subjects had the choice of expending memory effort (learning the locations of different colored blocks) or perceptual-motor effort (using keystrokes and mouse movements to uncover different areas of the display). A subject's total effort is the sum of the perceptual motor effort and the memory storage and recall effort.

The subjects were split into three groups; one group had to expend a low effort to uncover the grayed out areas, the second acted as a control, and the third had to expend a high effort to uncover the grayed out areas. The results showed that the subjects who had to expend a high perceptual-motor effort, uncovered grayed out area fewer times than the other two groups. These subjects also spent longer looking at the areas uncovered, and moved more colored blocks between uncoverings. The subjects faced with a high perceptual-motor effort reduced their total effort by investing in memory effort. Another consequence of this switch of effort investment, to use of memory, was an increase in errors made.

incentives improve cognitive performance...

Cognitive load theory...

What are the physical processes that generate the feeling of mental effort? The processes involved remain poorly understood.<sup>1241</sup> Proposals include: metabolic constraints (the brain

accounts for around 20% of heart output, and between 20% to 25% of oxygen and glucose requirements), but the energy consumption of the visual areas of the brain while watching television are higher than the consumption levels of those parts of the brain associated with difficult thinking, that the body's reaction to concentrated thinking is to try to conserve energy, for use in other opportunities that could arise in the immediate future, by creating a sense of effort.<sup>796</sup>

### 2.8.3 Personality & intelligence

Perhaps the most well-known personality test is the *Meyer-Briggs type indicator*, or MBTI (both registered trademarks of Consulting Psychologists Press). A lot has been published about the reliability and validity of this test,<sup>915,1094</sup> commercial considerations have also made it research difficult. The International Personality Item Pool (IPIP)<sup>510</sup> is a public domain measure containing over three thousand items that is growing in popularity.

Tests measuring something called IQ have existed for over 100 years. The results are traditionally encapsulated in a single number, but tests claiming to measure the various components of intelligence are available;<sup>327</sup> the model is that people possess a general intelligence  $g$  plus various domain specific intelligences, e.g., in tests involving maths the results would depend on  $g$  and maths specific intelligence and in tests involving use of language the results would depend on  $g$  and language specific intelligence.

The five factor theory of personality is a popular basis for personality and intelligence tests... it has its critics<sup>291</sup>...

Some personality traits are likely to be beneficial and others detrimental in those involved in some development activities... Without reliable techniques for measuring these characteristics...

Perhaps one of the most important traits is a willingness to spend large amounts of time engaged in a single activity...

emailed asking for data, may be on its way... ??

The results of personality and intelligence tests are sometimes included in the job selection process and sometimes in studies investigating developer mental characteristics.

Cultural intelligence hypothesis, specific set of social skills for exchanging knowledge in social groups children and chimpanzees have similar cognitive skills for dealing with the physical world, but children have more sophisticated skills for dealing with the social world.<sup>608</sup>

### 2.8.4 Attention

Most of the sensory information received by the brain does not need conscious attention and is handled by the unconscious. Conscious attention is like a spotlight shining cognitive resources on a chosen area. In today's world, there is often significantly more information available to a person than they have available attention resources, WEIRD people live in an attention economy.

People can direct attention to their internal thought processes and memories. Read the bold print in the following paragraph:

**Somewhere Among hidden the in most the spectacular Rocky Mountains cognitive near abilities Central City is Colorado the an ability old to miner select hid one a message box from of another. gold. We Although do several this hundred by people focusing have our looked attention for on it, certain they cues have such not as found type it style.**

What do you remember from the non-bold text? Being able to make a decision to direct conscious attention to inputs matching a given pattern is a technique for making good use of limited cognitive resources.

Much of the psychology research on attention has investigated how inputs from our various senses are handled. It is known that they operate in parallel and at some point there is a serial bottleneck, beyond which point it is not possible to continue processing input stimuli in parallel. The point at which this bottleneck occurs is a continuing subject of debate; there are early selection theories, late selection theories, and theories that combine the two.<sup>1059</sup>

A study by Rogers and Monsell<sup>1172</sup> used the two tasks of classifying a letter as a consonant or vowel, and classifying a digit as odd or even. The subjects were split into three groups. One group was given the letter classification task, the second group the digit classification task, and the third group had to alternate (various combinations were used) between letter and digit classification. The results showed that having to alternate tasks slowed the response times by 200 to 250 ms and the error rates went up from 2% to 3% to 6.5% to 7.5%. A study by Altmann<sup>28</sup> found that when the new task shared many features in common with the previous task (e.g., switching from classifying numbers as odd or even, to classifying them as less than or greater than five) the memories for the related tasks interfered, causing a reduction in subject reaction time and an increase in error rate.

and this relates to software engineering...

### 2.8.5 Risk taking

Making a decision based on incomplete or uncertain information involves an element of risk. How do people integrate risk into their decision-making process?

The term *risk asymmetry* refers to the fact that people have been found to be *risk averse* when deciding between alternatives that have a positive outcome, but are *risk seeking* when deciding between alternatives that have a negative outcome.<sup>vii</sup>

While there is a widespread perception that women are more risk averse than men, existing studies are either not conclusive or show a small effect<sup>431</sup> (many suffer from small sample sizes and dependencies on the features of the task subjects are given). For the case of financial risks the evidence<sup>239</sup> that men are more willing to take financial risks than women is more clear cut. The evidence from attempts to improve road safety is that "protecting motorists from the consequences of bad driving encourages bad driving".<sup>6</sup>

A study by Jones<sup>704</sup> investigated the possibility that some subjects in an experiment involving recalling information about previously seen assignment statements were less willing to risk giving an answer when they had opportunity to specify that in real-life they would refer back to previously read code. Previous studies<sup>700, 701</sup> had found that a small percentage of subjects consistently gave much higher rates of "would refer back". One explanation is that these subjects had a smaller short term memory capacity than other subjects (STM capacity does vary between people), another explanation is that these subjects are much more risk averse than the other subjects.

The Domain-Specific Risk-Taking (DOSPERT) questionnaire<sup>149, 1435</sup> was used to measure subject's risk attitude. The results showed no correlation between risk attitude (as measured by DOSPERT) and number of "would refer back" responses.

### 2.8.6 Decision-making

Human decision-making often takes place in an environment of incomplete information and limited decision-making resources (e.g., working memory capacity and thinking time); people have been found to adopt various strategies to handle this situation,<sup>894</sup> balancing the predicted cognitive effort required to use a particular decision-making strategy against the likely accuracy achieved by that strategy.

The term *bounded rationality*<sup>1257</sup> is used to describe an approach to problem solving applied when limited cognitive resources are available. A growing number of studies<sup>497</sup> have found that the methods used by people to make decisions and solve problems are often close enough to optimal<sup>viii</sup>, given the resources available to them; even when dealing with financial matters.<sup>919</sup> If people handle money matters in this fashion, their approach to software development decisions is unlikely to fare any better.

?

A so-called *irregular choice* occurs when a person who chooses from the set of items {A, B, C} does not choose B from the set {A, B}; irregular decision makers have been found<sup>1344</sup> to

---

<sup>vii</sup> While studies<sup>604</sup> based on subjects drawn from non-WEIRD societies sometimes produce different results, this book assumed developers are WEIRD.

<sup>viii</sup> Some researchers interpret bounded rationality as human decision-making producing results as-if people optimize under cognitive constraints, while others<sup>493</sup> don't require people to optimize, but to use algorithms that produce results that are good enough.

be more common among younger (18–25) subjects, but less are common with older (60–75) subjects.

Consumer research into understanding how shoppers decide among competing products uncovered a set of mechanisms that are applicable to decision-making in general, e.g., decision-making around the question of which soap will wash the whitest is no different from the question of whether an **if** statement or a **switch** statement be used.

Before a decision can be made, a decision-making strategy has to be selected. People have been found to use a number of different decision-making strategies and this section discusses some of these strategies and the circumstances under which people might apply them; Payne, Bettman, and Johnson<sup>1065</sup> covers the topic in detail.

A developer who has an hour to write a program knows there is not enough time to make complicated trade-offs among alternatives. There is a hierarchy of responses for how people deal with time pressure:<sup>1065</sup> they work faster; if that fails, they may focus on a subset of the issues; if that fails, they may change strategies (e.g., from alternative based to attribute based).

Decision strategies differ in several ways, for instance, some make trade-offs among the attributes of the alternatives (making it possible for an alternative with several good attributes to be selected instead of the alternative whose only worthwhile attribute is excellent); they also differ in the amount of information that needs to be obtained and the amount of cognitive processing needed to make a decision.

**The weighted additive rule:** this requires the greatest amount of effort, but delivers the most accurate result (strong attachments to a particular point of view has been found to influence the weight given to evidence that contradicts this view<sup>1347</sup>). The steps are as follows:

- build a list of attributes for each alternative,
- assign a value to each of these attributes,
- assign a weight to each of these attributes (these weights could, for instance, reflect the relative importance of that attribute to the person making the decision, or the probability of that attribute occurring),
- for each alternative, sum the product of each of its attributes' value and weight,
- select the alternative with the highest sum.

When dividing two numbers in a loop, a developer has to decide whether it is worthwhile testing the denominator against zero before the division. The obvious attributes are performance and reliability. A comparison would decrease performance, an attribute whose weight will depend on the time-critical nature of the loop. The benefit from making a comparison is the potential to improve reliability by detecting an error case.

**The equal weight heuristic:** this is a simplification of the weighted additive rule that assigns the same weight to every attribute. This heuristic might be applied when accurate information on the importance of each attribute is not available and a decision is made to use equal weights.

**The frequency of good and bad features heuristic:** estimating a numerical measure for an attribute may be very difficult and a binary measurement based on good/bad is sometimes good enough (i.e., looking at things in black and white). This heuristic is a simplification of the equal weight heuristic.

**The majority of confirming dimensions heuristic:** estimating an absolute measure for an attribute may be very difficult, but a yes/no answer can be given to the question: “Is the value of attribute X greater (or less) for alternative A compared to alternative B?”. The answer can be used to determine which alternative has the most (or least) of each attribute.

The algorithm starts by selecting a pair of alternatives, compare each matching attribute of the alternatives and select the alternative that has the greater number of winning attributes, the winning alternative is then paired with an unpaired alternative and the compare/select steps are repeated until one of the alternatives is left.

**The satisficing heuristic:** the result of this heuristic can depend on the order in which alternatives are checked and often does not check all alternatives. The steps are as follows:

- assign a cutoff, or aspirational, level that must be met by each attribute,
- for each alternative:
  - check each of its attributes against the cutoff level, rejecting the alternative if the attribute is below the cutoff,
  - if there are no attributes below the cutoff value, accept this alternative,
- if no alternative is accepted, revise the cutoff levels associated with the attributes and repeat the previous step.

When selecting a library function to obtain some information, the list of attributes might include the amount of information returned and the format it is returned in (relative to the format it is required to be in). Once a library function meeting the developer's minimum aspirational level is found, additional effort need not be invested in finding a better alternative.

**The lexicographic heuristic:** has a low effort cost, but it might not be very accurate. It can also be intransitive; with X preferred to Y, Y preferred to Z, and Z preferred to X. The steps are as follows:

- determine the most important attribute of all the alternatives,
- find the alternative that has the best value for the selected most important attribute,
- if two or more alternatives have the same value, select the next most important attribute and repeat the previous step using the set of alternatives whose attribute values tied,
- the result is the alternative having the best value on the final, most important, attribute selected.

An example of the intransitivity that can occur, when using this heuristic, might occur when writing software for embedded applications. Here the code has to fit within storage units that are available in fixed-size increments (e.g., 8 K chips). It may be possible to increase the speed of execution of an application by writing code for special cases, or have generalized code that is more compact, but slower. Table 2.8 shows some commonly seen, alternatives:

Alternative	Storage Needed	Speed of Execution
X	7 K	Low
Y	15 K	High
Z	10 K	Medium

Table 2.8: Storage/Execution performance alternatives.

Based on storage minimization, X is preferred to Y. Because storage comes in 8 K increments there is no preference, based on the storage attribute, between Y and Z. Based on speed of execution, Y is preferred to Z, and Z is preferred to X.

**The habitual heuristic:** looks for a match of the current situation against past situations, it does not contain any evaluation function (although there are related heuristics that evaluate the outcome of previous decisions). The single step for this heuristic is: select the alternative chosen last time for that situation.

Going with a winning solution suggests one of two possibilities:

- so little is known that once a winning solution is found, it is better to stick with it than to pay the cost (time and the possibility of failure) of looking for a better solution that might not exist,
- the developer has previously performed an extensive analysis and the best solution is known.

**Copying others:** Find good solutions to problems is hard, doing what others do can be a cost effective strategy. *When Strategies:* copy when established behavior is unproductive, copy when asocial learning is costly, copy when uncertain. *Who Strategies:* copy the majority, copy if rare, copy successful individuals, copy if better, copy if dissatisfied, copy good social learners, copy kin, copy "friends", copy older individuals.

Studies have found that having to justify a decision can affect the choice of decision-making strategy used.<sup>1346</sup> One strategy that handles accountability is to select the alternative that the perspective audience is thought most likely to select.<sup>1345</sup> People who have difficulty determining which alternative has the greatest utility tend to select the alternative that supported the best overall reasons (for choosing it).<sup>1259</sup>

Requiring developers to justify why they have not followed existing practice can be a two-edged sword. Developers can respond by deciding to blindly follow everybody else (a path of least resistance), or they can invest effort in evaluating alternatives (not necessarily cost effective behavior, since the invested effort may not be warranted by the expected benefits). The extent to which some people will blindly obey authority was chillingly demonstrated in a number of studies by Milgram.<sup>944</sup>

Social pressure can cause people to go along with the decision voiced by those currently in the room. A study by Asch<sup>57</sup> asked a group of seven to nine subjects to individually state which of three black strips they considered to be the longest (see Figure 2.45), the group sat together in front of the stripes subjects and could interact with each other; all the subjects, except one, were part of the experiment and in 12 of 18 questions selected a stripe that was clearly not the longest (i.e., the majority gave an answer clearly at odds with visible reality). It was arranged that the actual subject did not have to give an answer until hearing the answers of most of the other subjects.

The actual subject in 24% of groups always gave the correct answer, in 27% of groups the subject agreed with the incorrect majority answer between eight and twelve times and just under half varied in the extent to which they followed the majority decision. When the majority selected the most extreme incorrect answer, i.e., the shortest stripe, subjects giving an incorrect answer selected the less extreme incorrect answer in 20% of cases.

Social conformity as a signal of belonging... Corporate IT fashion Figure 1.13...

Informational cascades<sup>139</sup>...

Without any evidence about the efficiency of existing practice, it is not unreasonable to assume that the practice arose out of a random event that became dominant. A more efficient way of performing the same practice may not become established because it does not provide enough benefit to overcome existing practices...

How a problem is posed can have a large impact on the decision made.

A study by Regnell, Höst, och Dag, Beremark and Hjelm<sup>1150</sup> asked 10 subjects to assign a relative priority to two lists of requirements (subjects' had a total budget of 100,000 units and had to assign units to requirements). One list specified that subjects should prioritize the 17 high level requirements and the other list specified that a more detailed response, in the form of prioritizing every feature contained within each high level requirement, be given.

Comparing the totals for the 17 high level requirements (summing the responses for the detailed list), showed that the correlation was not very strong (the mean across 10 subjects was 0.46, see [projects/prioritization.R](#)).

A study by Hofman<sup>627</sup> investigated... [developers/Hofman-exp1.R](#)

### 2.8.6.1 Expected utility and Prospect theory

The outcome of events is often uncertain. If events are known to occur with probability  $p_i$ , with each producing a value of  $X_i$ , then the expected outcome value is given by:

$$E[x] = p_1X_1 + p_2X_2 + p_3X_3 + \dots + p_nX_n$$

For instance, given a 60% chance of winning £10 and a 40% chance of winning £20, the expected winnings are:  $0.6 \cdot 10 + 0.4 \cdot 20 = 14$

When comparing the costs and benefits of an action, decision makers often include information on their current wealth, e.g., a bet offering a 50% chance of winning £1 million and a

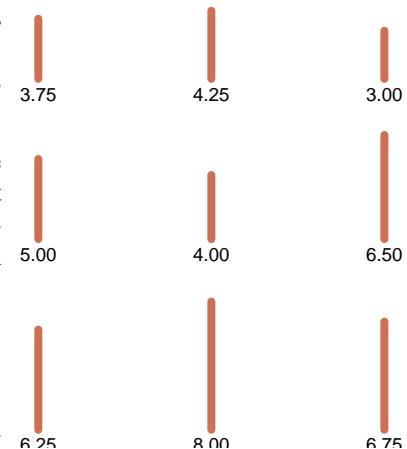


Figure 2.45: Each row shows a scaled version of the three stripes, along with actual lengths in inches, from which subjects were asked to select the longest. Based on Asch,<sup>57</sup> code

50% chance of losing £0.9 million has an expected utility of £0.05 million; would you take this bet unless you were very wealthy?

£20 might not be considered to be worth twice as much as £10. The mapping of an action's costs and benefits to a decision maker's particular circumstances is made by what is known as a *utility function*, the above equation becomes (where  $W$  is the decision maker's current wealth):

$$E[x] = p_1u(W + X_1) + p_2u(W + X_2) + p_3u(W + X_3) + \dots + p_nu(W + X_n)$$

In some situations a decision maker's current wealth might be effectively zero, e.g., they have forgotten their wallet, or because they have no authority to spend company money on work related decisions (personal wealth is unlikely to factor into many work decisions).

Figure 2.46 shows possible perceived utilities of an increase in wealth. A risk neutral decision maker perceives the utility of an increase in wealth as being proportional to the increase (green,  $u(w) = w$ ), a risk loving decision maker perceives the utility of an increase in wealth as being proportionally greater than the actual increase (red,  $u(w) = w^2$ ), while a risk averse decision maker perceives the utility of an increase having proportionally less utility (blue,  $u(w) = \sqrt{w}$ ).

A study by Kina, Tsunoda, Hata, Tamada and Igaki<sup>754</sup> investigated the decisions made by subjects (20 open source developers) when given a choice between two tools, each having various probabilities of providing various benefits, e.g., *Tool*<sub>1</sub> which always saves 2 hours of work, or *Tool*<sub>2</sub> which saves 5 hours of work with 50% probability, and has no effect on work time with 50% probability.

Given a linear utility function, *Tool*<sub>1</sub> has an expected utility of 2 hours, while *Tool*<sub>2</sub> has an expected utility of 2.5 hours. The results showed 65% of developers choosing *Tool*<sub>1</sub>, and 35% choosing *Tool*<sub>2</sub>: clearly those 65% were not using a linear utility function; use of a square-root utility function would produce the result seen:  $1 \cdot \sqrt{2} > 0.5 \cdot \sqrt{5} + 0.5 \cdot \sqrt{0}$ .

A study by Stewart, Reimers and Harris<sup>1315</sup> ...

Prospect theory...

### 2.8.6.2 Overconfidence

Overconfidence is a false belief which can create unrealistic expectations, and lead to hazardous decisions being made (e.g., to allocate insufficient resources to complete a job).

There is evidence from simulation studies<sup>691</sup> that there are benefits to being overconfident in some situations, averaged over a population (see [developers/overconfidence/0909.R](#)). A study<sup>58</sup> of commercialization of new inventions found that while inventors are significantly more confident and optimistic than the general population, the likely return on their investment of time and money in their invention is negative; a few provide a huge payoff.

A study by Lichtenstein and Fishhoff<sup>843</sup> asked subjects general knowledge questions, with the questions divided into two groups, hard and easy. Figure 2.47 shows that subjects' overestimated their ability (x-axis) to correctly answer hard questions, but underestimated their ability to answer easy questions; solid line denotes perfect self-knowledge.

These, and subsequent results, show that the skills and knowledge that constitute competence in a particular domain are the same skills needed to evaluate one's (and other people's) competence in that domain. *Metacognition* is the term used to denote the ability of a person to accurately judge how well they are performing.

Peoples' belief in their own approach to getting things done can result in them ignoring higher performing alternatives;<sup>49</sup> this behavior has become known as *the illusion of control*.<sup>812</sup>

It might be thought that people who have previously performed some operation would be in a position to make accurate predictions about future performance on those operations. However, studies have found<sup>1181</sup> that while people do base predictions of future duration on their memories of how long past events took, these memories are systematic underestimates of past duration. People appear to underestimate future event duration because they underestimate past event duration.

### 2.8.6.3 Time discounting

People seek to consume pleasurable experiences sooner and to delay their appointment with painful experiences, i.e., people tend to accept less satisfaction in the short-term than could be obtained by pursuing a longer-term course of action. Studies have found that animals, including humans, use a hyperbolic discount function for time variable preferences.<sup>340</sup> The hyperbolic delay discount function is:

$$v_d = \frac{V}{1+kd}$$

where:  $v_d$  is the delayed discount,  $V$  the undiscounted value,  $d$  the delay and  $k$  some constant.

A property of this hyperbolic function is that curves with different values of  $V$  and  $d$  can cross; see Figure 2.48. In this example, the perceived present value of two future rewards (red and blue lines) starts with red being initially greater than blue, as time passes (i.e., the present time moves right and the delay receiving the rewards decreases) the reward denoted by the blue line out-grows the red line, until there is a reversal in perceived present value. When both rewards are far in the future, the larger amount has the greater perceived value, studies have found<sup>760</sup> that subjects switch to giving a higher value to the lesser amount as the time of receiving the reward gets closer.

a variety of models have been proposed<sup>361</sup> ...

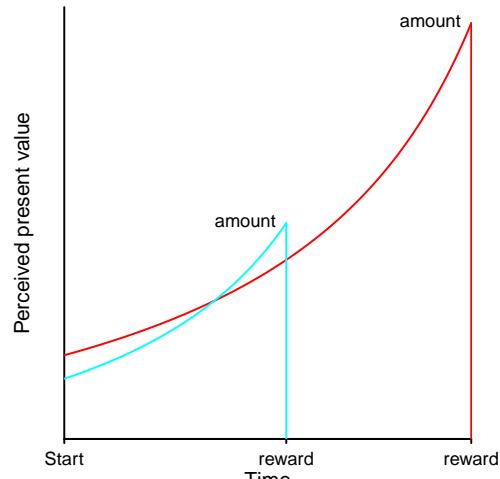


Figure 2.48: Perceived present value (moving through time to the right) of two future rewards. [code](#)

### 2.8.7 Personal integrity

Reusing other peoples' code...?

?

The unacknowledge use of third-party software Section 5.3.1...

### 2.8.8 Miscellaneous characteristics

There are a variety of human characteristics that probably have some impact on the software development process, but the size of the impact... including the following:

- Reaction time has an ex-Gaussian distribution... emailed for data...
- Fitts' law: ...?
- Hick's law: Time taken to decide which item, from a known list of items, is currently visible<sup>589</sup>...
- Ageing effects: The Seattle Longitudinal Study<sup>1205</sup> has been following the intellectual development of over six thousand people since 1956 (surveys of the individuals in the study are carried out every seven years). Findings include: "... there is no uniform pattern of age-related changes across all intellectual abilities, ..." and "... reliable replicable average age decrements in psychometric abilities do not occur prior to age 60, but that such reliable decrements can be found for all abilities by age 74 ..." An analysis<sup>158</sup> of the workers on the production line at a Mercedes-Benz assembly plant found that productivity did not decline until at least up to age 60.

Years of experience vs. experimental performance....

Changes in a person's social and economic standing over time are likely to have a much larger impact than changes in mental ability. For instance, commitments outside work (e.g., family) reduce the time available for acquiring new skills; people who are paid more (an individual's salary is likely to increase over time) are expected to deliver more (people promoted to their maximum level of incompetence; it's cheaper to hire younger people who are already familiar with new technologies than retrain more expensive older people). and the data backing up these claims...

Figure 8.12 shows an age distribution of developers.

Altruism... Cooperation...?

## 2.9 Developer performance

Companies want to hire those people who will give the best software development performance. Currently the only reliable way of evaluating developer performance is by measuring developer outputs (a reasonably accurate model of the workings of human mental operations remains in the future).

One operational characteristic of the brain that can be estimated is the number of operations potentially performed per second (a commonly used method of estimating the performance of silicon-based processors).

The brain might simply be a very large neural net, so there may be no instructions to count as such; Merkle<sup>935</sup> used the following approaches to estimate the number of synaptic operations per second (the supply of energy needed to fire neurons limits the number that can be simultaneously active, in a local region, to between 1% and 4% of the neurons in that region<sup>835</sup>):

- multiplying the number of synapses ( $10^{15}$ ) by their speed of operation (about 10 impulses/second) gives  $10^{16}$  synapse operations per second (if the necessary energy could be delivered to all of them at the same time),
- the retina of the eye performs an estimated  $10^{10}$  analog add operations per second. The brain contains  $10^2$  to  $10^4$  times as many nerve cells as the retina, suggesting that it can perform  $10^{12}$  to  $10^{14}$  operations per second,
- a total brain power dissipation of 25 watts (an estimated 10 watts of useful work) and an estimated energy consumption of  $5 \cdot 10^{-15}$  joules for the switching of a nerve cell membrane provides an upper limit of  $2 \cdot 10^{15}$  operations per second.

A synapse switching on and off is the same as a transistor switching on and off in that they both need to be connected to other switches to create a larger functional unit. It is not known how many synapses are used to create functional units, or even what those functional units might be. The distance between synapses is approximately 1 mm and sending a signal from one part of the brain to another part requires many synaptic operations, for instance, to travel from the front to the rear of the brain requires at least 100 synaptic operations to propagate the signal. So the number of synaptic operations per high-level, functional operation is likely to be high. Silicon-based processors can contain millions of transistors; the potential number of transistor-switching operations per second might be greater than  $10^{14}$ , but the number of instructions executed is significantly smaller.

Although there have been studies of the information-processing capacity of the brain (e.g., visual attention<sup>1399</sup> and storage rate into long-term memory<sup>168,806</sup>), we are a long way from being able to deduce the likely work rates of the components of the brain used during code comprehension.

Processing units need a continuous energy to function. The brain does not contain any tissue that stores energy and obtains all its energy needs through the breakdown of blood-borne glucose. Consuming a glucose drink has been found to increase blood glucose levels and enhance performance on various cognitive tasks.<sup>743</sup> Also, fluctuations in glucose levels have an impact on an individual's ability to exert self-control,<sup>470</sup> with some glucose intolerant individuals not always acting in socially acceptable ways.<sup>ix</sup>

Do people vary in the ability to supply energy to the brain and remove waste products?...

How do developers differ in their measurable output performance?

Although much talked about, there has been little research on individual developer productivity. Claims of a 28-to-1 productivity difference between developers is sometimes still bandied about. The, so called *Grant-Sackman study*<sup>538</sup> is based on an incorrect interpretation of a summary of the data.<sup>1114</sup> The data shows a performance difference of around 6-to-1 between developers using different systems for creating software (i.e., batch vs. online).

---

<sup>ix</sup> There is a belief in software development circles that consumption of chocolate enhances cognitive function. A systematic review of published studies<sup>1209</sup> found around a dozen papers addressing this topic, with three finding some cognitive benefits and five finding some improvement in mood state.

Most organizations do not attempt to measure the mental characteristics of developer job applicants; unlike many other jobs for which individual performance is an important consideration. Whether this is because of the existing non-measurement culture, lack of reliable measuring procedures, or fear of frightening off prospective employees is not known.

A study of development and maintenance costs of programs written in C and Ada<sup>1481</sup> found no correlation between salary grade (or employee rating) and rate of bug fix or feature implementation rate.

One metric used in software testing is number of faults found. In practice non-failing tests, written by software testers, are useful because they provide evidence that particular functionality behaves as expected.

A study by Iivonen<sup>660</sup> analysed the defect detection performance of those involved in testing software at several companies. Table 2.9 shows the number of defects detected by six testers (all but the first column show percentages), along with self-classification of the seriousness, followed by the default status assigned by others.

Tester	Defects	Extra Hot	Hot	Normal	Open	Fixed	No fix	Duplicate	Cannot reproduce
A	74	4	1	95	12	62	26	12	0
B	73	0	56	44	15	87	6	2	5
C	70	0	29	71	36	71	24	0	4
D	51	0	27	73	33	85	6	0	9
E	50	2	16	82	30	89	9	0	3
F	18	0	22	78	22	64	14	0	21

Table 2.9: Defects detected by six testers (some part-time and one who left the company during the study period) and their status. Data from Iivonen.<sup>660</sup>

A performance comparison, based on defects reported, requires combining these figures (and perhaps others, e.g., likelihood of being experienced by a customer) into a value that can be reliably compared across testers. Is the weight assigned to defects classified as “No fix” is larger than that given to “Cannot reproduce” or “Duplicate”?

To what extent would a tester’s performance, based on measurements involving one software system in one company be transferable to another system in the same company or another company? Iivonen interviewed those involved in testing to find out what characteristics were thought important in a tester. Knowledge of customer processes and use cases was a common answer; this usage knowledge enables testers to concentrate on those parts of the software that customers are most likely to use and be impacted by incorrect operation, it also provides the information needed to narrow down the space of possible input values.

Knowledge of the customer ecosystem and software development skills are the two blades that have to mesh together to create useful software systems.

## 2.10 Biased behavior

What environmental characteristics might have driven the known cognitive biases, i.e., why might people give answers that are not consistent with those returned by the use of mathematical logic?

*Anchoring* is a cognitive bias where people assign too much weight to the first piece of information they obtain, relating to the topic at hand. A study by Jørgensen and Sjøberg<sup>720</sup> asked professionals and students to estimate the development effort for a project, with one group of subjects being given a low estimate from the *customer* another a low estimate and the third group no customer estimate. The results (see [developer/anchor-estimate.R](#)) showed that estimates from subjects given a high/low customer estimate were much higher/lower than subjects who did not receive any customer estimate.

A study by Jørgensen and Grimstad<sup>715</sup> asked subjects to estimate the number of lines of code they wrote per hour, with subjects randomly split into two groups; one anchored with the question: “Did you on average write more or less than 1 Line of Code per work-hours in your last project?” and the other with: “Did you on average write more or less than 200 Lines of Code per work-hours in your last project?” Fitting a regression model to the results showed the form of the question changed the mean estimate by around 72 lines (sd 10)... [developers/estimation-biases.R](#)

Software application change over time and users' opinions of them are also likely to change. Does the release of a bug ridden version of an application create anchor a negative opinion, even after releases that fix the bugs? If the number of faults experienced in an application slowly increases over time, does user opinion remain anchored at earlier, bug free levels?

A person exhibits *confirmation bias* if they tend to interpret ambiguous evidence as (incorrectly) confirming their current beliefs about the world. For instance, developers interpreting program behavior as supporting their theory of how it operates, or using the faults exhibited by a program to conform their view that it was poorly written...

confirmation bias leads to overconfidence (people believing in some statement, on average, more strongly than they should)...

When shown data from a set of observations, a person might propose a set of rules that the processes generating the data adhere to. Given the opportunity to test proposed rules, what strategy are people likely to use?

A study by Wason,<sup>1430</sup> which became known as the 2–4–6 Task, asked subjects to discover a rule known to the experimenter; subjects' guessed a rule, told it to the experimenter, who told them whether the answer was correct. For instance, on being informed that the sequence 2–4–6 was an instance of the experimenter's rule, possible subject rules might be "two added each time" or "even numbers in order of magnitude", when perhaps the actual rule was "three numbers in increasing order of magnitude".

An analysis of the rules created by subjects found that most were test cases designed to confirm a hypothesis about the rule (known as a *positive test strategy*), with few test cases attempting to disconfirm a hypothesis. Some subjects declared rules that were mathematically equivalent variations of rules they had already declared.

The use of a positive test strategy was seen as a deficiency because of the influential work of Popper,<sup>1103</sup> a philosopher, who proposed that scientists should perform experiments designed to disprove their hypothesis. Studies<sup>1199</sup> of the hypothesis testing strategies used by scientists show that positive testing is the dominant approach.

An analysis by Klayman and Ha<sup>767</sup> investigated the structure of the problem subjects were asked to solve. The problem is a search problem, find the experiment's rule, and in some environments a positive test strategy is more effective for solving search problems, compared to a negative test strategy.

A positive test strategy is more effective when the sought after rule describes a minority case, i.e., there are more cases not covered by the rule, or when the hypothesized rule includes roughly as many cases as the actual rule, i.e., the hypothesized rule is about the right size. Klayman and Ha claimed these conditions hold for many real-world rule search problems and is therefore an adaptive strategy; the real-worldness claim continues to be debated.<sup>996</sup>

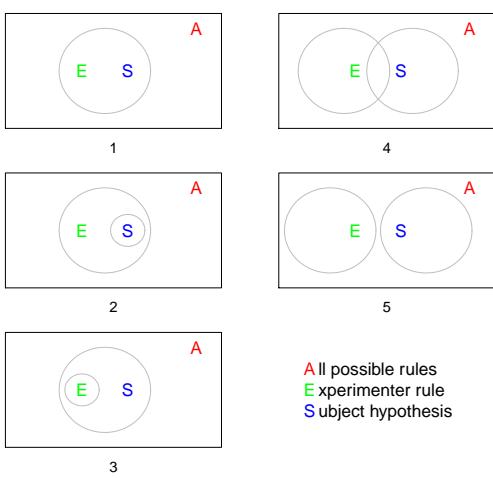


Figure 2.49: The five possible ways in which experimenter's rule and subject's rule hypothesis can overlap, in the space of all possible rules; based on Klayman et al.<sup>767</sup> code

# Chapter 3

## Cognitive capitalism

### 3.1 Introduction

Software systems are intangible goods that are a product of cognitive capitalism. Many of the existing capitalist structures are oriented towards the production of tangible goods, and are slowly adapting to handle intangible goods.<sup>126,581</sup>

Economic analysis is the primary tool used, in this chapter, to discuss cognitive capitalism. Economics deals with tradeable quantities, such as money, time and pleasure.

The analysis in this book is oriented towards the producers of software rather than its consumers, consequently this chapter seeks to focus on maximization of return on investment for producers. Much of the current software engineering research agenda has been set by a few large organizations that have actively engaged the research community, e.g., the U.S. Department of Defence and NASA. As customers of software systems these organizations have promoted a customer orientated focus to software economics, i.e., minimizing customer costs and risks, with little interest in vendor profitability and risk.

Some of the cognitive biases that influence peoples' decision-making are discussed in Chapter 2, covering human cognitive characteristics, e.g., people tend to be risk averse for positive outcomes and risk seeking for negative outcomes.

What percentage of their income do software companies spend on developing software? A study by Mulford and Misra<sup>976</sup> of 100 companies in Standard Industry Classifications (SIC) 7371 and 7372<sup>i</sup>, with revenues exceeding \$100 million during 2014–2015, found that total software development costs were around 13% of revenue; see Figure 3.1. Marketing, profit...

The financial package...

#### 3.1.1 Some definitions

This section briefly covers a few commonly encountered economic concepts.

The term *cost/benefit* applies when making a decision about whether to invest or not, while the term *cost-effectiveness* applies when a resource is available and has to be used wisely or when an objective has to be achieved as cheaply as possible.

*Return on investment* (ROI) is defined as:

$$R_{est} = \frac{B_{est} - C_{est}}{C_{est}}$$

where:  $B_{est}$  is the estimated benefit and  $C_{est}$  the estimated cost.

Both the cost and the benefit estimates are likely to contain some amount of uncertainty and the minimum and maximum ROI are given by:

$$R_{est} - \delta R = \frac{B_{est} - \delta B}{C_{est} + \delta C} - 1, \text{ and } R_{est} + \delta R = \frac{B_{est} + \delta B}{C_{est} - \delta C} - 1$$

where:  $\delta$  is the uncertainty in the corresponding quantity.

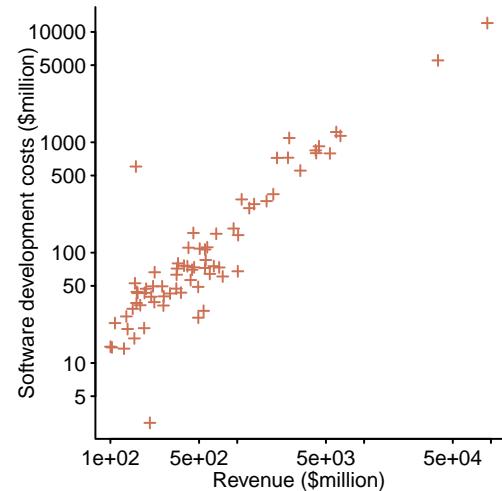
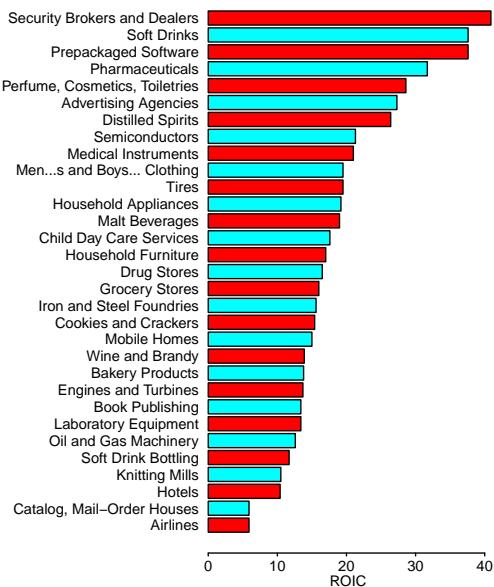


Figure 3.1: Company revenue (\$millions) against total software development costs. Data from Mulford et al.<sup>976</sup> code

<sup>i</sup> Computer programming services and Prepackaged software.



The ROI uncertainty is unlikely to have an extreme value and its expected value is given by:<sup>161</sup>

$$E[\delta R] \approx \frac{B_{est}}{C_{est}} \sqrt{\left(\frac{\delta B}{B_{est}}\right)^2 + \left(\frac{\delta C}{C_{est}}\right)^2}$$

When a resource is invested, the opportunity to use it to make an alternative investment, possibly with a greater return, is lost; this loss of opportunity is known as an *opportunity cost*. For instance, when deciding between paying for Amazon spot instances<sup>124</sup> at a rate similar to everybody else and investing time trying to figure out an algorithm that makes it possible to successfully bid at much lower prices, the time spent figuring out the algorithm is an opportunity cost (i.e., the time spent is a lost opportunity for doing something else, which may have been more profitable).<sup>ii</sup>

Moral hazard...

Cash-flow...

## 3.2 Investment decisions

A pound/dollar in the hand today is worth more than a pound/dollar in the hand tomorrow. The return required by a high risk investment is greater than the return required from a low risk investment.

Figure 3.3 shows the development cost of video games (where the cost was more than \$50million). The high risk of a market that requires a large upfront investment, to create a new product for an uncertain return, is offset by the possibility of a high return.

### 3.2.1 Discounting for time

A dollar today is worth more than a dollar tomorrow, because today's dollar can be invested and earn interest; by tomorrow the amount will have increased in value. The present value (*PV*) of a future payoff, *C*, can be calculated from:

$$PV = \text{discount\_factor} \cdot C$$

where *discount\_factor* < 1 and usually represented by: *discount\_factor* =  $(1 + r)^{-1}$ , where *r* is known as the *rate of return* (also known as the *discount rate* or the *opportunity cost of capital*) and represents the size of the reward demanded by investors for accepting a delayed payment; it is often quoted over a period of a year.

The *PV* over *n* years (or whatever period *r* is expressed in) is given by:

$$PV = \frac{C}{(1 + r)^n}$$

When comparing multiple options, expressing their cost in terms of present value enables them to be compared on an equal footing.

The *internal rate of return* (also known as the *discounted cash-flow rate of return*, *profitability index*, and *interest-rate-of-return method*) is a widely used method for calculating the profitability of an investment. It is the value of *r* for which the following is true:

$$\sum_{n=1}^L \frac{\text{Cash\_flow\_in\_year } n}{(1 + r)^n} = 0$$

where *L* is the lifetime of a system in years.

Take as an example the choice of between spending \$250,000 purchasing a test tool, or the same amount on hiring testers; assuming the tool will make an immediate cost saving of \$500,000 (by automating various test procedures), while hiring testers will result in a saving of \$750,000 in two years time. Which is the better investment (assuming a 10% discount rate)?

$$PV_{tool} = \frac{\$500,000}{(1 + 0.10)^0} = \$500,000$$

<sup>ii</sup> There is also the risk that the result of successfully reverse engineering the pricing algorithm results in Amazon changing the algorithm.<sup>124</sup>

$$PV_{testers} = \frac{\$750,000}{(1 + 0.10)^2} = \$619,835$$

Based on these calculations, hiring the testers is the better option, i.e., it has the greater present value.

During the development of a system the cashflow will be negative (i.e., more money is being spent than is coming in from customers). Once completed the income from customer sales has to cover the original, and any ongoing, development costs.

### 3.2.2 Taking risk into account

The calculation in the previous section did not take risk into account. What if the tool did not perform as expected, what if some testers were not as productive as hoped? A more realistic calculation of present value takes into account the risk of future payoffs not being as large as expected.

A risky future payoff is not worth as much as a certain future payoff for the same amount invested. The risk is factored into the discount rate to create an *effective discount rate*:  $k = r + \theta$  (where  $r$  is the risk-free rate <sup>iii</sup> and  $\theta$  a premium that depends on the amount of risk). The formulae for present value becomes:

$$PV = \frac{C}{(1+k)^n}$$

Recognizing that both  $r$  and  $\theta$  can vary over time we get:

$$PV = \sum_{i=1}^t \frac{return_i}{1+k_i}$$

where  $return_i$  is the return during period  $i$ .

Repeating the preceding example, assuming a 15% risk premium for the testers option, we get:

$$PV_{tool} = \frac{\$500,000}{(1+0.10)^0} = \$500,000$$

$$PV_{testers} = \frac{\$750,000}{(1+0.10+0.15)^2} = \$480,000$$

Taking an estimated risk into account suggests that buying the tool is the better option.

The benefits that have been compared both require an investment to be made. Comparing investment costs requires that calculating the values at a given moment in time; the following calculates today's cost:

Purchasing the tool is a one time, up front, payment:

$$investment\_cost_{tool} = \$250,000$$

The cost of the testers approach is likely to be dominated by monthly salary costs. If the testing cost is \$10,416.67 per month for 24 months, the total cost after two years, in today's terms, is (a 10% annual interest rate is close to 0.8% per month):

$$investment\_cost_{testers} = \sum_{m=0}^{23} \frac{\$10,416.67}{(1+0.008)^m} = \$10,416.67 \left[ \frac{1 - (1+0.008)^{-22}}{1 - (1+0.008)^{-1}} \right] = \$211,042.90$$

Spending \$250,000 over two years is equivalent to spending \$211,042.90 today. Investing \$211,042.90 at 10% today, provides a fund that supports spending \$10,416.67 per month for 24 months.

This calculation assumed a constant testing effort; Chapter 5 discusses the distribution of various development activities over the lifetime of a development project.

*Net Present Value* (NPV) is defined as:

$$NPV = PV - investment\_cost$$

Plugging in the calculated values gives:

---

<sup>iii</sup> The rate offered by government bonds maturing within the same time frame are commonly used as a value for the risk-free rate.

$$NPV_{tool} = \$500,000 - \$250,000 = \$250,000$$

$$NPV_{testers} = \$480,000 - \$211,042.90 = \$268,957.10$$

Based on NPV, hiring testers is the better option.

Alternatives to NPV, their advantages and disadvantages, are discussed by Brealey<sup>176</sup> and Raffo.<sup>1140</sup> One commonly encountered rule in rapidly changing environments is the payback rule, which requires that the investment costs of a project be recovered within a specified period; the *payback period* is the amount of time needed to recover investment costs (a shorter payback period being preferred to a longer one).

Estimating using Monti-Carlo simulation...

A cost/benefit analysis of the decision of whether to fix a known fault now, has to take into account the probability that it will never be experienced by a customer (see Figure 11.75) and the possibility of reputational damage if it results in a major loss for important customers.

Accurate estimates for the value of different options require good estimates for the discount rate and the impact of risk. The discount rate represents the risk-free element and the closest thing to a risk-free investment is government bonds and securities. Information on these rates are freely available. Governments face something of a circularity problem in how they calculate the discount rate for their own investments. The US government discusses these issues in its “Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs”<sup>1444</sup> and at the time of this writing the revised Appendix C specified rates varied between 0.9% over a three-year period and 2.7% over 30 years. Commercial companies invariably have to pay higher rates of interest than the US Government.

Analyzing software development risk is a very hard problem. Information on previous projects carried out within the company can offer some guidance on the likelihood of developers meeting productivity targets. In a broader context the market conditions also need to be taken into account, for instance: how likely is it that other companies will bring out competing products? Will demand for the application still be there once development is complete?

The way in which options are used to control the risks associated with buying shares on the stockmarket (e.g., portfolio analysis<sup>427</sup> using the Black-Scholes equation<sup>iv</sup>) suggests a parallel with the options and risks associated with developing software. However, any parallel is superficial because several fundamental components of portfolio theory, as applied to financial instruments such as stocks and securities, don't apply to software development; differences include:

- liquidity is necessary to enable the composition of a portfolio to be changed, at will, by investing or divesting (i.e., buying or selling). Software is not a liquid asset, once an investment has been made in writing code it is very difficult to immediately divest a percentage of the investment in this code (i.e., it is a sunk cost), or to make further investments simply for the purpose of maintaining the risk composition of a portfolio,
- historical information on the performance of a stock is an essential input to portfolio risk calculations. Historical data is rarely available on one, let alone all, of the major performance factors involved in software development.

While portfolio theory is not directly applicable to software development, it may be applicable in other computer related contexts, such as company wide IT decisions...

### 3.2.3 Incremental investments and benefits

When investments and benefits occur at intervals, the value calculations can be involved. The following example is based on the idea that it is possible to make an investment when writing or maintaining code that produces a benefit, i.e., reduces subsequent maintenance costs.

As a minimum, any investment in reducing subsequent maintenance costs must be recouped in subsequent maintenance.

---

<sup>iv</sup> The derivation of this formula assumes: the price of the underlying asset follows a lognormal random walk, investors can continually adjust their position at no cost, the risk-free rate is known and the underlying asset does not pay dividends (i.e., a pure investment that never reaches a stage where it generates revenue).

Let  $d$  be the original development cost,  $m$  the base yearly maintenance cost and  $r$  the interest rate, we start by keeping things simple and assume  $m$  is the same for every year of maintenance; the total amount paid over  $y$  years is:

$$\text{Total\_cost} = d + \sum_{m=1}^y \frac{m}{(1+r)^m} = m \left[ \frac{1 - (1+r)^{-(y+1)}}{1 - (1+r)^{-1}} - 1 \right] \approx d + y \cdot m$$

with the approximation applying when  $r$  is small.

If we make an investment of  $i\%$  for all implementation work, in expectation of achieving a  $b\%$  reduction in maintenance effort during each year, we get:

$$\text{Total\_cost} = d \cdot (1+i) + y \cdot m \cdot (1+i) \cdot (1-b)$$

For this investment to break-even the following condition must hold:

$$d \cdot (1+i) + y \cdot m \cdot (1+i) \cdot (1-b) < d + y \cdot m$$

expanding and simplifying (where  $ym$  can be replaced by the exact expression given earlier) gives a lower bound for the ratio  $\frac{b}{i}$ :

$$1 + \frac{d}{ym} < \frac{b}{i} + b$$

In practice many systems are replaced after a surprisingly short period. What relationship must the ratio  $\frac{b}{i}$  have for an investment to be worthwhile, taking into account system survival rate (i.e., the risk that there is no future maintenance)?

Let  $s$  be the probability that a system will survive each year, total system cost is now:

$$\text{Total\_cost} = d + M \cdot s + M \cdot s^2 + M \cdot s^3 + \dots + M \cdot s^y$$

where:  $M = m \cdot (1+i) \cdot (1-b)$ . This can be simplified to:

$$\text{Total\_cost} = d + M \frac{s(1-s^y)}{1-s}$$

and the break-even relationship is now:

$$1 + \frac{d}{M s(1-s^y)} < \frac{b}{i} + b$$

The development/maintenance break-even ratio now depends on the yearly maintenance multiplied by a factor that depends on the system survival rate, and not the total maintenance cost.

An analysis of system lifespan data in the Section 4.6.2 finds:  $s \rightarrow 0.88$  and  $\frac{d}{M} \rightarrow 0.56$  and the above requirement becomes:  $1 + 0.56 \cdot 0.29 = 1.16 < \frac{b}{i} + b$ . The relative percentage of yearly maintenance effort is large enough to be considered continuous development and the percentage reduction in maintenance does not have to be much larger than the investment.

This analysis only considers systems that have been delivered and deployed; projects are sometimes cancelled before reaching this stage and including these in the analysis would increase the benefit/investment break-even ratio.

While a surprisingly large percentage<sup>374</sup> of program features may never be used, which features will not be used is unknown at the time of implementation...

Most files are only ever edited by one person, emailed and promised data...<sup>1340</sup>

### 3.2.4 Quality adjusted process

Products may evolve over time, with new features being added and existing features removed. Techniques for calculating quality adjusted prices have been developed... Figure 3.8<sup>1325</sup>...

### 3.2.5 Information asymmetry

In an information economy those with access to the applicable information have a significant advantage over those who do not. In markets where no information about product quality is available, low quality drives out higher quality.<sup>16</sup>

The term *information asymmetry* describes the situation where one party in a negotiation has access to important, applicable, information that is not available to the other parties. The information is important in the sense that it can be used to make more accurate decisions.

The development group that created a software system has the luxury of not needing to underestimate the cost of work on the next revision, to get the business. The customer is aware of the group's intimate familiarity with the workings of the software and has greater reason to believe their estimates, or to fear the uncertainty of firing an alternative vendor. Unless the quoted price is exorbitant, the customer will be unwilling to take the risk of having the new development performed by a group quoting a lower price.

Figure 3.4 shows the ratio of actual to estimated hours of effort involved in 25 releases of one application over six years (release 1.1 to 9.1, figures for release 1 are not available).<sup>649</sup> Possible reasons for the estimated effort being so much higher than the actual effort include: cautiousness on the part of the supplier and a willingness to postpone implementation of features planned for the current release to a future release (one of the benefits of trust built up between supplier and customer in an ongoing relationship is flexibility of scheduling).

### 3.3 Company economics

Possible barriers to companies entering an existing market include:

- Economies of scale (a supply side effect): producing a product in large volumes allows firms to spread their fixed costs over more units, improving efficiency (requires that production be scalable in a way that allows existing facilities to produce more). Competing against a firm producing at volume requires large a capital investment by a company entering the market, to prevent them being at a cost disadvantage.

A large percentage of the cost of software production is spent on people, who provide few opportunities for economies of scale (there may even be diseconomies of scale, e.g., communication overhead increases with head count),

- Network effects (a demand-side effect): customers willingness to buy from a supplier increases with the number of existing customers. Buyers more likely to trust a larger company. Competing against a firm selling into a market with network effects requires large a capital investment by a company entering the market, to offer incentives so that a significant percentage of customers will switch suppliers...
- Switching costs:<sup>413</sup> a *switching cost* is an investment specific to the current supplier that must be duplicated for a new supplier, e.g., retraining staff and changes to procedures, e.g., the UK government paid £37.6 million transition costs to the winning bidder of the ASPIRE contract, with £5.7 million paid to the previous contract holder to facilitate changeover.<sup>279</sup>

Information asymmetry can create switching costs by deterring competition; to encourage competition in bidding on the replacement ASPIRE contract, the incumbent had inside knowledge and was perceived to be strong, the UK government contributed £8.6 million towards bidders' costs.<sup>279</sup>

Switching costs can incentivise vendors to concentrate on servicing existing customers, rather than investing effort acquiring new customers (who buy from other vendors).<sup>414</sup>

Ho hum... the data...

The competitive forces faced by a company include:<sup>1105</sup> rivalry between existing competitors, bargaining power of suppliers, bargaining power of buyers, possibility of new entrants, and possibility product being substituted by alternative products or services...

The Red Queen model,<sup>94</sup> data promised...

The value of a business has two components: tangible goods such as buildings, equipment and working capital, and intangible assets which are a product of knowledge (e.g., employee know-how, intellectual property and customer switching costs)... Figure 1.8...

An investment in the creation of intangible good is a sunk cost; if the project fails there is unlikely to be a secondary market where what has been created can be sold...

Governments have been relatively slow to include intangible investments in the calculation of GDP (1999 in the US and 2001 in the UK). The accounting treatment of intangibles depends on whether it is purchased from outside the company, requiring it to be treated as an asset, or generated internally where is often treated as an expense<sup>976</sup>... A study by Hulten<sup>651</sup> investigated Microsoft's intangible capital.

Sources of funding... technology driving economic cycles<sup>1076</sup>...

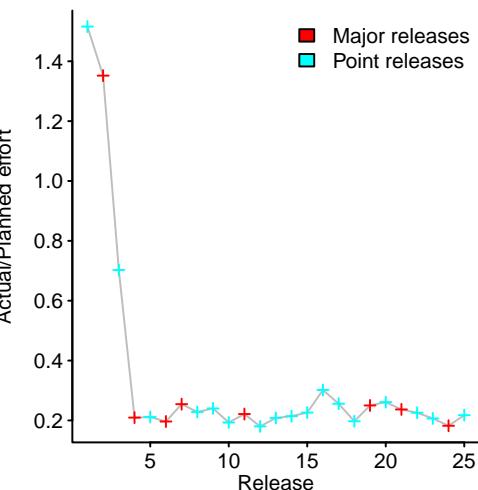


Figure 3.4: Ratio of actual to estimated hours of effort to enhance an existing product, for 25 versions of one application. Data from Huijgens et al.<sup>649</sup> code

### 3.3.1 Cost accounting

The purpose of cost accounting is to help management make efficient costing decisions. In traditional industries the primary inputs are the cost of raw materials (e.g., the money needed to build the materials needed to build a widget) and labor costs (e.g., the money paid to the people involved in converting the raw materials into a widget); other costs might include renting space where people can work and the cost of consumables such as electricity.

The production of software systems is people driven and people are the primary cost source.

How much does a software developer cost?

The direct cost of a developer is the money they are paid, plus any taxes levied by the government, on an employer, for employing somebody (e.g., national insurance contributions in the UK <sup>v</sup>).

Developers have to work somewhere, they need computers, coffee, tables and chairs and a whole host of items are used during the production of a software system; these overheads, or expenses, are indirect costs. In a company that employs more than one person, who develops software, may be necessary to have rules for deciding how indirect costs are divided up across individuals, product groups or multiple-sites...

Rules may be driven by the expediency of getting products written, or mandated by management driven by the need for the accounts to ??? Which/whose budgets do costs come out of???

Activity-based costing (ABC) is a system of assigning costs for a product or service based on the activities required. This system provides detailed information about the cost of providing every product and service, which coupled with the corresponding income enables ROIs to be calculated. Implementing this system requires that every employee keep a complete record of what they spend their time doing.

Cost centers...

The fully loaded cost of an engineer...

Revenue per employee... Meti data... Does not include contractors, consultants...

Incremental cost of fixing one bug is minimal vs. calculating the cost of a by dividing total support costs by the number of bugs...

### 3.3.2 The shape of money

Money is divided up and categorized in various ways for a variety of different reasons. Figure 3.5 shows the way in the UK and US tax authorities require registered companies to apportion their income to various cost centers.

Within a company a group of senior people take the money available and allocate a budget for each of the organizational groups within the company. These groups, in turn, divide up their available budget to their own organizations groups and so on. This discrete allocation of funds can have an impact on how the costs involved in software development are calculated.

Take the example of a development project where testing is broken down into two phases, i.e., integrating testing and acceptance testing, each having its own budget, say  $B_i$  and  $B_a$ .

One technique sometimes used for measuring the cost of faults is to divide the cost of finding them (i.e., the allocated budget) by the number of faults found. For instance, if 100 faults are found during integration testing, the cost per fault in this phase is  $\frac{B_i}{100}$ ; if five faults are found during acceptance testing, the cost per fault is  $\frac{B_a}{5}$ .

One way of reducing the cost per fault found during acceptance testing would be to reduce the effectiveness of integration testing. For a fixed budget the cost per fault decreases when the number of faults increases.

This accountancy approach to measuring the cost of faults gives the appearance that it is more costly to find faults later in the process, compared to faults found earlier.<sup>159</sup> A conclusion created by the artifact of a fixed budget and a smaller percentage of faults found later in the development process.

<sup>v</sup> At the time of writing this was zero-rated up to a threshold, then 12% of employee earnings, increasing on reaching an upper threshold.

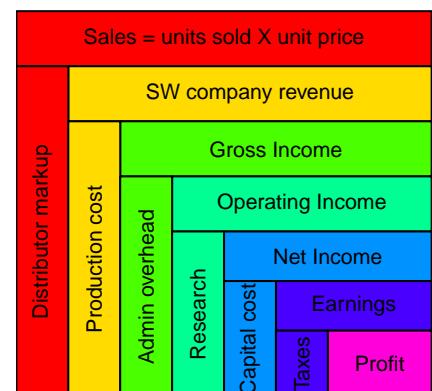


Figure 3.5: Accounting practice for breaking down income from sales... [code](#)

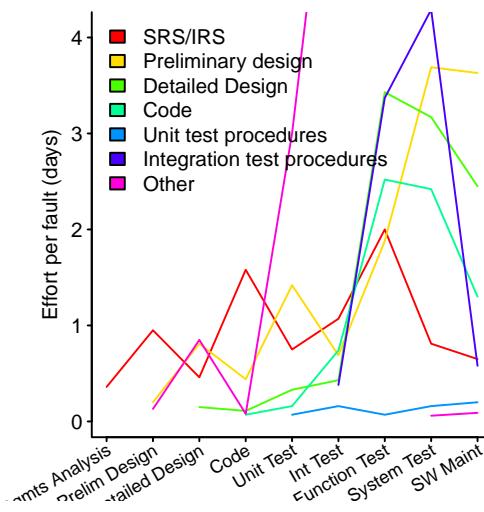


Figure 3.6: Average effort (in days) used to fix a defect detected in a given phase (x-axis) that had been introduced in an earlier phase (colored lines), introduced in an earlier phase (total of 38,120 defects in projects at Hughes Aircraft). Data extracted from Willis et al.<sup>1457</sup> [code](#)

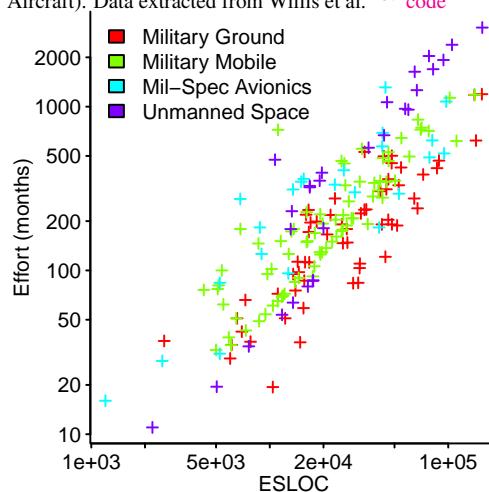


Figure 3.7: Months of developer effort needed to produce systems containing a given number of lines of code, for various application domains. Data from Gayek et al.<sup>482</sup> [code](#)

Time taken to fix faults is less likely to be influenced by accounting artifacts (but does not include the cost of the resources used). Figure 3.6 shows the average number of days used to fix a defect detected in a given phase (x-axis) that had been introduced in an earlier phase (colored lines), based on 38,120 defects in projects at Hughes Aircraft.<sup>1457</sup>

### 3.3.3 Valuing software

Like any other item, if nobody is willing to pay money for the software it has zero sales value. The opportunity cost of writing the software from scratch, along with the uncertainty of being able to complete the task, is what enables existing software to be worth more than the original cost of creating it.

Software might be valued in terms of the cost of producing it. When the production cost is not known, the amount of code it contains (often measured in lines) is sometimes suggested as a proxy, even though customers pay for a useful system to be produced and the amount of code involved is of little interest to them.

A study by Gayek, Long, Bell, Hsu and Larson<sup>482</sup> obtained general effort/size information on 452 military/space projects. Figure 3.7 shows the lines of code and the developer effort (in months) used to create them. There is an order of magnitude variation in developer effort (i.e., costs) for the same ESLOC and the ESLOC vs. Effort relationship e appears to vary between end-user application domains.

Commercial companies are required to keep accurate financial accounts. The purpose of these accounts is to provide essential information for those with a financial interest in the company, including governments seeking to tax profits. Official accounting organizations have created extensive, and ever-changing, rules for producing company accounts, including methods for valuing software and other products of intellectual effort.<sup>597</sup> In the US, the Financial Accounting Standards Board has issued an accounting standard “FASB Codification 985-20” (FAS 80<sup>433</sup> was used until 2009) covering the “Accounting for the Costs of Computer Software to Be Sold, Leased, or Otherwise Marketed”. This allows software development costs to be treated either as an expense or capitalized (i.e., treated like the purchase of an item of hardware). An expense is tax-deductible in the financial year in which it occurs, but the software does not appear in the company accounts as having any value; the value of a capitalized item is written down over time (i.e., a percentage of the value is tax-deductible over a period of years), but has a value in the company accounts.<sup>753</sup>

The decision on how software development costs appear in the company accounts can be driven by the desire to project a certain image to interested outsiders (e.g., the company is worth a lot because it owns valuable assets<sup>3</sup>) or to minimize tax liabilities. A study by Mulford and Roberts<sup>977</sup> of 207 companies (primarily industry classification SIC 7372) in 2006 found that 30% capitalized some portion of their software development, while a study by Mulford and Misra<sup>976</sup> of 100 companies in 2015 found 18% capitalizing their development.

Software made available under an Open Source license may be available for free, but this does not mean it has no financial value...

Value based on expected lifetime... lifetime of SQL tables discussed in Section 4.8.4..

Wiederhold<sup>1449</sup> gives an overview of value software and any associated intellectual property...

## 3.4 Maximizing profit

Why do companies employ people to write software?

Incentives, money, reputation<sup>644</sup>...

Focus on the most profitable markets...

One study<sup>747</sup> of Android Apps found that 80% of reviews were made by people owning a subset of the available devices (approximately 33%). Given the cost of testing an App in the diverse Android ecosystem it is cost effective to target the subset with the largest customer base.

Taxation issues... purchase vs leasing... contract vs employees The tax system of many countries treats software as an asset... Arranging so that income is held offshore to reduce tax liabilities<sup>1197</sup>...

### 3.4.1 Product/service pricing

Vendors can set whatever price they like for their product or service. The ideal is to set the price that maximises profit; maximum profit can be zero or negative, i.e., there is no guarantee that sales will cover the cost of development and the sales operation. Even given extensive knowledge of potential customers and the competitors, setting prices is very difficult. Shy<sup>1251</sup> provides a detailed analysis of pricing issues.

Product pricing is a complicated subject<sup>991</sup>...

Like everything else, software products are worth what customers are willing to pay. When prices quoted by vendors are calculated on an individual basis, the inputs to the formula can be viewed as a metric for how much money the customer might be able pay.<sup>1295</sup>

Customer expectation based on related existing products acts as a useful ballpark to start the estimation process; Figure 3.8 illustrates the relative performance pricing used by Intel. It may be possible to charge more for a product that provides more advantages, or perhaps the price has to match that of the established market leader and these advantages have to be used to convince existing customers to switch.

Figure 3.16 shows the prices charged by several established C/C++ compiler vendors. In 1986<sup>vi</sup> Zorland entered the market with a £29.95 C compiler, an order of magnitude less than what most other vendors were charging at the time. This price was low enough for many developers to purchase a copy out of company petty cash and Zorland C quickly gained a large customer base. Once the product established a quality reputation Zorland management were able to increase prices to be the same as those charged by other major vendors (Zorland sounds very similar to the name of another major vendor of the time, Borland; letters from company lawyers are hard evidence that somebody in authority thinks you are worth investing money on; Zorland became Zortech).<sup>vii</sup>

Many companies delegate to managers the authority to purchase low cost items, with the numeric value of low price increasing with seniority. The upper bound on the maximum price that can be charged for a product or service that can be relatively quickly sold to businesses is the purchasing authority of the target decision maker. Once the cost of an item reaches some internally specified value (perhaps a few thousand pounds or dollars) companies require a more bureaucratic purchase process to be followed, perhaps involving a purchasing committee or even the company board. Navigating a company's bureaucratic processes requires a sales rep and a relatively large investment of time, increasing the cost of sales and requiring a significant increase in selling price; a price dead-zone exists between the maximum amount managers can independently sign-off on and the minimum amount it is worth selling via sales reps.

Supply and demand is a commonly cited economic pricing model. The supply curve is the quantity of an item that a supplier is willing and able to supply (i.e., sell) to customers at a given price, and the demand curve quantity of a product that will be in demand (i.e., bought by customers) at a given price. If these two curves intersect, the intersection point gives the price/quantity at which suppliers and customers are willing to do business; see Figure 3.9.

Events can occur that cause either curve to shift. For instance, the cost of manufacturing an item may increase/decrease shifting the supply curve up/down on the price axis; or customers may find a cheaper alternative, shifting the demand curve down on the price axis.

In some established markets enough historic information has accumulated for reasonably accurate supply/demand curves to be drawn. Predicting the impact of changing circumstances on supply-demand curves remains a black art for many products and services.

Software is a relatively new market and one that continues to change relatively quickly. This makes estimating supply/demand curves for software products little more than wishful thinking.

Product pricing has an effect on the volume of products sold...

Other pricing issues include site licensing, discounting and selling through third-parties.<sup>1237</sup> However, the lack of data prevents these issues being discussed here.

<sup>vi</sup> Richard Stallman's email announcing the availability of the first public release of gcc was sent on 22 March 1987.

<sup>vii</sup> Being in the compiler business your author had copies of all the major compilers, and Zorland C was the compiler of choice for a year or two. Other low price C compiler vendors were unable to increase their prices because of quality issues relative to other products on the market, e.g., Mix C.

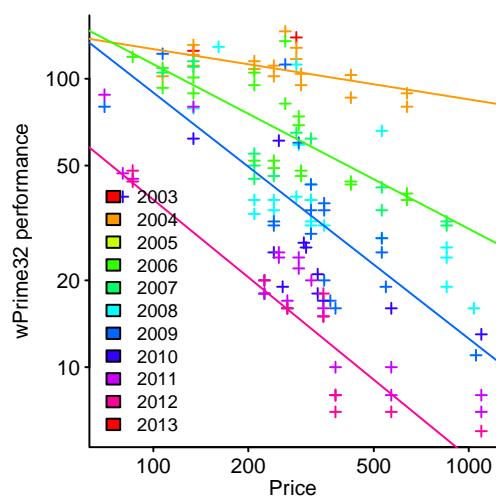


Figure 3.8: Introductory price and performance (measured using wPrime32 benchmark) of various Intel processors between 2003-2013. Data from Sun.<sup>1325</sup> code

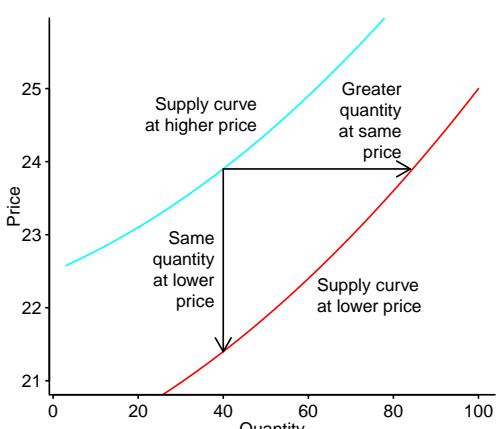
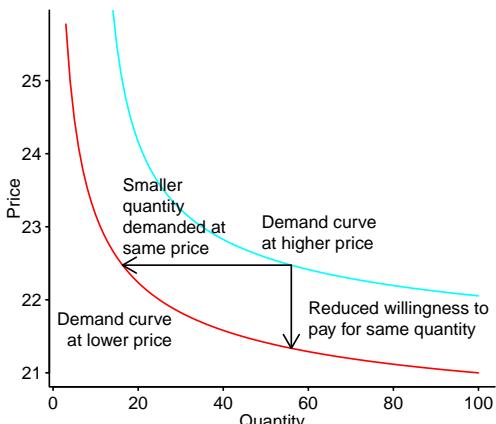


Figure 3.9: Example supply and demand curves. code

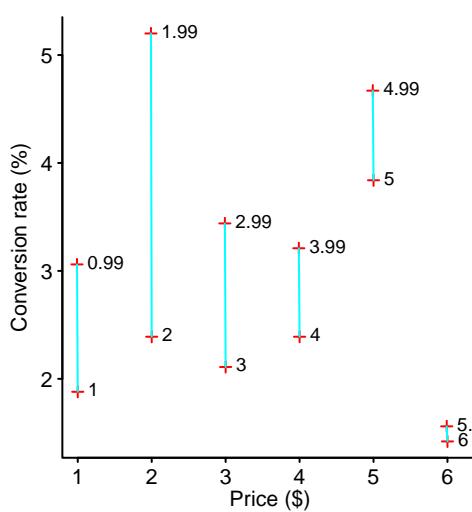


Figure 3.10: Rates at which product sales are made on Gumroad at various prices; lines join prices that differ in 1¢, e.g., \$1.99 and \$2. Data from Nichols.<sup>1360</sup> code

Listed prices are often slightly below a round value, e.g., £3.99 or £3.95 rather than £4.00. People have been found to perceive this kind of price difference to be greater than the actual numerical difference<sup>1352</sup> (the value of the left digit and the numeric distance effect have been used to explain this behavior). Figure 3.10 shows the impact of visually distinct, small price differences, have on the rate of sale of items listed on Gumroad (a direct to consumer sales website). Other consumer price perception effects include precise prices vs. round prices<sup>1353</sup> ...

The price of a basket of common products over time is used to calculate the consumer price index and changes in product prices over time can be used to check the accuracy of official figures.<sup>1054</sup>

### 3.4.2 Predicting sales volume

A critical economic question for any system intended to be sold to multiple customers is, how many systems are likely to be sold.

When one product substitutes another, new for old, market share of the new product is well fitted by a logistic equation<sup>439</sup> whose maximum is the size of the existing market.

Estimating the likely sales volume for a product intended to fill a previously unmet customer need, or one that is not a pure substitution, is extremely difficult (if not impossible).

The Bass model<sup>102</sup> addresses customer adoption of a new product (the following analysis deals with the case where customers are likely to make a single purchase). The model divides customers into innovators, people who are willing to try something new, and imitators, people who buy products they have seen other people using it (this is a diffusion process and the model belongs to the class of diffusion models). The interaction between innovators, imitators and product adoption, at time  $t$ , is given by the following relationship:

$$\frac{f(t)}{1-F(t)} = p + qF(t)$$

where:  $F(t)$  is the fraction of the total of those who will eventually adopt (i.e., purchased) by time  $t$ ,  $f(t)$  is the probability of purchase at time  $t$  (i.e., the derivative of  $F(t)$ ,  $f(t) = dF(t)/dt$ ),  $p$  the probability of a purchase by an innovator (known as the *coefficient of innovation*), and  $q$  the probability of a purchase by an imitator (known as the *coefficient of imitation*).

This non-linear differential equation<sup>viii</sup> has the following exact solution for the cumulative number of adopters (up to time  $t$ ):

$$F(t) = \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p}e^{-(p+q)t}}$$

from which the instantaneous number of adopters (at time  $t$ ) can be obtained:

$$f(t) = \frac{(p+q)^2}{p} \frac{e^{-(p+q)t}}{\left[1 + \frac{q}{p}e^{-(p+q)t}\right]^2}$$

Actual sales, up to, or at, time  $t$ , are calculated by multiplying by  $m$ , the total number of product adopters.

When innovators dominate (i.e.,  $q \leq p$ ), sales decline from an initial high-point; when imitators dominate (i.e.,  $q > p$ ), peak sales of  $m\left(\frac{1}{2} - \frac{p}{2q}\right)$  occurs at time  $\frac{m}{p+q} \log \frac{q}{p}$ , before declining; the expected time to adoption is:  $E(T) = \frac{m}{q} \log \frac{p+q}{p}$ .

The exact solution applies to a continuous equation, but in practice sales data is discrete (e.g., monthly, quarterly, yearly). In the original formulation the model was reworked in terms of a discrete equation and solved using linear regression (to obtain estimates for  $p, q$ ); however, this approach produces biased results. Benchmarking various techniques<sup>1070</sup> finds that fitting the above non-linear equation to the discrete data produces the least biased results.

Vendors want good estimates of likely sales volume as early in the sales process as possible, but building accurate models requires data covering a non-trivial range of the explanatory

variable (time in this case). Figure 3.11 shows the number of Github users during its first 58 months and Bass models fitted to the first 24, 36, 48 and 58 months of data.<sup>ix</sup>.

The Bass model includes just two out of the many possible variables that could affect sales volume, models that include more variables have been developed<sup>956</sup> and Intel have used an extended Bass Model to improve forecasting of design wins.<sup>1468</sup>

The Bass model can be extended to handle successive, overlapping generations of a product; the following example is for two generations:<sup>1019</sup>

$$S_1(t) = F_1(t)m_1 - F_2(t - \tau_2)F_1(t)m_1 = F_1(t)m_1(1 - F_2(t - \tau_2))$$

$$S_2(t) = F_2(t - \tau_2)(m_2 + F_1(t)m_1)$$

where:  $S_i(t)$  are all sales up to time  $t$  for product generation  $i$ ,  $m_i$  the total number who adopt generation  $i$  and  $\tau_2$  the time when the second generation can be bought;  $p_i$  and  $q_i$  are the corresponding purchase probabilities for each generation.

The Bass equation is just one of the components of a model aiming to predict product sales, other factors that need to be taken into account include advertising spend and variability in market size caused by price changes. Monte Carlo simulation can be used to model the interaction of the various components.<sup>1388</sup>

Repeat sales...?

How much value, as perceived by the customer, does each major component add to a system? One technique used to answer this question is Hedonic regression, which fits a regression model to data on product price and configuration data. Economists use this approach in the calculation of the consumer price index. A study by Stengos and Zacharias<sup>1306</sup> performed by hedonic analysis of the Personal Computer market, based on data such as price, date, CPU frequency, hard disc size, amount of RAM, screen width and presence of a CD; see [economics/0211\\_Computers.R](#) for details.

The ease with which software can be copied makes piracy an important commercial issue. Studies<sup>501, 1432</sup> have extended the Bass model to include terms for the influence that users of pirated software have on causing others to purchase a copy of the software (e.g., Word processors and Spreadsheet programs between 1987 and 1992). The results, based on the assumptions made by the models and the data used, suggested that around 85% of users run pirated copies; see [economics/MPRA/](#). The Business Software Alliance calculates piracy rates by comparing the volume of software sales against an estimate of the number of computers in use, a method that has a high degree of uncertainty because of the many assumptions involved<sup>1098</sup> (see [economics/piracyICSS – 2010.R](#)).

Figure 3.12 shows how software sales volume lags behind sales of the hardware needed to run it. A study by Marchand<sup>893</sup> investigated factors that might have an impact on console game software sales.

In some ecosystems, e.g., mobile, many applications are only used for a short period after they have been installed; see Figure 4.34.

In some markets most sales are closed just before the end of each yearly quarter, e.g., enterprise software. A study by Larkin<sup>816</sup> investigated the impact of non-linear incentive schemes<sup>x</sup> on the timing of deals closed by salespeople whose primary income came from commission on the sales they closed. These accelerated commission schemes create an incentive for salespeople to book all sales in a single quarter.

Figure 3.13 shows the number of deals closed by week of the quarter and the average agreed discount. Reasons for the significant peak in the number of deals closed at the end of the quarter include salespeople gaming the system to maximise commission and customers holding out for a better deal. Larkin concluded that the peak was salesperson driven.

Mobile sales revenue?

<sup>viii</sup> It has the form of a Riccati equation.

<sup>ix</sup> Chapter 11 provides further evidence that predictions outside the range of data used to fit a model can be very unreliable.

<sup>x</sup> The percentage commission earned in a non-linear scheme depends on the total value of sales booked in the current quarter, increasing at specified points, e.g., a salesperson booking \$250,000 in a quarter earns 2% commission, while a salesperson booking over \$6 million earns a commission of 25%; that first \$250,000 earns the first salesperson a commission of \$5,000, while it earns the second salesperson \$62,500.

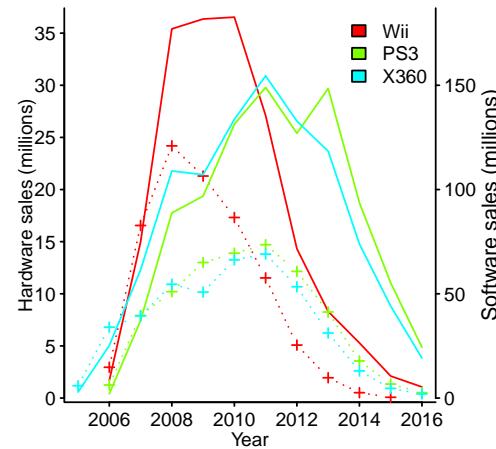


Figure 3.12: Sales of game software (solid lines) for the corresponding three major seventh generation hardware consoles (dotted lines). Data from VGChartz.<sup>1404</sup> [code](#)

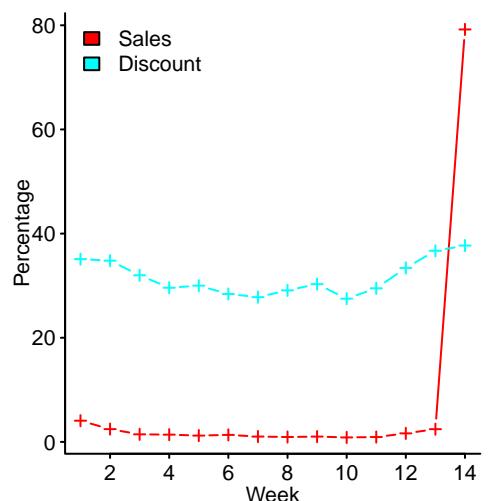


Figure 3.13: Percentage of sales closed in a given week of a quarter. Data from Larkin.<sup>816</sup> [code](#)

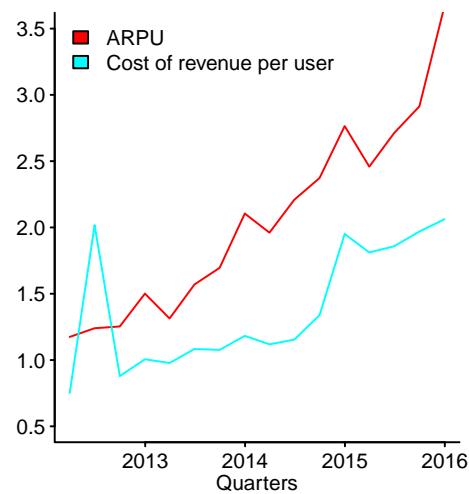


Figure 3.14: Facebook's ARPU and cost of revenue per user. Data from Facebook's 10-K filings.<sup>404,405</sup> [code](#)

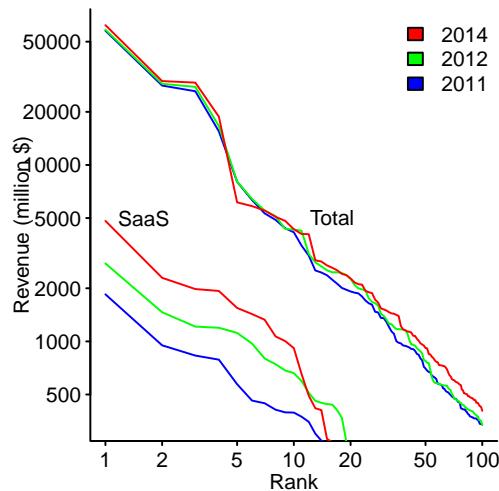


Figure 3.15: Top 100 software companies ranked by total revenue (in millions of dollars) and ranked by Software-as-a-Service revenue. Data from PwC.<sup>1127–1129</sup> [code](#)

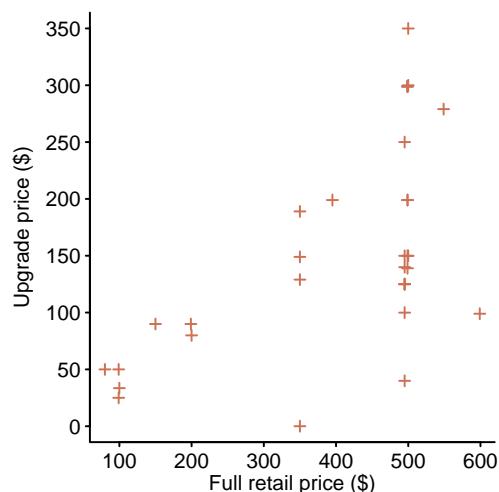


Figure 3.16: Various vendor's retail price and upgrade prices for C and C++ compilers available under MS-DOS and Microsoft Windows between 1987 and 1998. Data kindly provided by Viard.<sup>1405</sup> [code](#)

### 3.4.3 Managing customers as investments

Acquiring new customers can be very expensive and it is important to maximise the revenue from those that are acquired. What is the total value of a customer to a company?

A person who uses a product without paying for it may be valued as a user. Facebook values its users (whose eyeballs are the product that Facebook sells to its customers: advertisers) using ARPU, and defines it as "...total revenue in a given geography during a given quarter, divided by the average of the number of MAUs in the geography at the beginning and end of the quarter."<sup>xi</sup> Figure 3.14 shows ARPU and cost of revenue per user (difference is one definition of profit).

If the customer makes regular payments of  $m$ , the *customer lifetime value* (CLV) is given by (assuming the payment is made at the end of the period; simply add  $m$  if payment occurs at the start of the period):

$$CLV = \frac{mr}{(1+i)} + \frac{mr^2}{(1+i)^2} + \frac{mr^3}{(1+i)^3} + \dots$$

where:  $r$  is the customer retention rate for the period and  $i$  the interest rate for the period. If these payments occur for a maximum of  $n$  periods we get:

$$CLV = m \frac{r}{1-r+i} \left[ 1 - \left( \frac{r}{1+i} \right)^n \right]$$

which simplifies to the following, as  $n \rightarrow \infty$ :

$$CLV = m \frac{r}{1-r+i}$$

A year maintenance agreement is a common form of regular payment in the business world, another is the sale of new versions of the product to existing customers (often at a discount).

Every year existing customers have to see a worthwhile benefit in renewing their maintenance agreement. Worthwhile benefits include promises of new product features and continuing to fix problems encountered by the customer. Vendor's need to continually offer product improvements means it is not in their interest to include too many new features, or fix too many faults, in each release; something always has to be left for the next release. and no data... .

The customer retention rate,  $r$ , is the important quantity...

A study by Viard<sup>1405</sup> investigated software upgrades, in particular C and C++ compilers available under Microsoft DOS and Windows between 1987 and 1998... Figure 3.16...

Customer complaints...? Warrant claims on software sales... /usr1/data.../warranty-claims

### 3.4.4 Commons-based peer-production

Code distributed under an Open source license is often created using a non-contract form of production<sup>126</sup>?...

Motivation... .

What are the currencies of commons-based peer-production?...

## 3.5 Game theory

Trust is an essential component of economic activity, and is not always taken into account.<sup>319</sup>

Iterated prisoner's dilemma... Strategies that can handle noisy input... .

??

??,??

The term *vaporware* is used, in the software industry, to describe the practice of announcing products with a shipment date well in the future. These announcements are a signalling

<sup>xi</sup> ARPU—Average Revenue Per User, MAU—Monthly Average Users.

mechanism, whose uses include, keeping existing customer happy that a product has a future and deterring potential competitors;<sup>339</sup> there may or may not be a real product under development. A study by Bayus, Jain and Rao<sup>112</sup> investigated the delay between promised availability and actual availability of 123 software products. Figure 3.17 shows that the interval between preannouncement and promised availability date does not have much impact on the average delay between promised and actual delivery date.

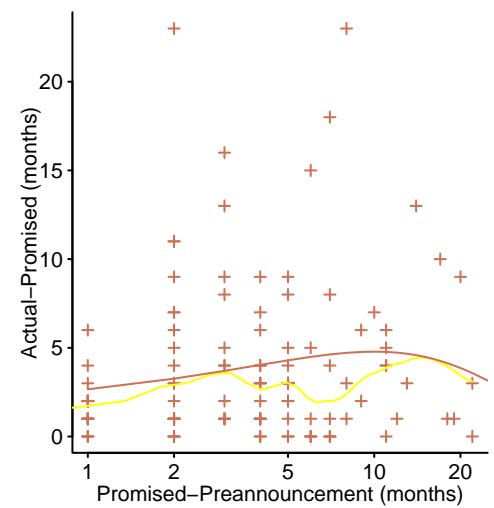


Figure 3.17: Interval between product preannouncement date and its promised availability date against delay between promised date and actual date product became available. Data from Bayus et al.<sup>112</sup> [code](#)



# Chapter 4

## Ecosystems

### 4.1 Introduction

Customer demand drives the creation of software ecosystems; once established they are driven by the economics of fix or replace.

The firestorm driving the continual displacement of existing systems has kept the focus of practice on development (of new systems and major enhancements to existing systems). Post-firestorm, the focus of practice in software engineering is an infrastructural discipline.<sup>1373</sup>

Customer demand for software systems takes many forms, including: a means of reducing costs, creating new products sold for profit, tools that people can use to enhance their daily life, experiencing the pleasure of creating something (e.g., writing open source)...

Customer demands cannot always be supplied at a profit and customer demand for new products introduced by companies is often insufficient for them to be profitable.<sup>i</sup>

Like the industrial revolution,<sup>21</sup> established industries wanting to reduce their labor costs created the initial customer demand for computer systems. The creation of new industries based around the use of computers came later. The first computers were built from components manufactured for other purposes (e.g., radio equipment), as the market for computers grew companies sprang up to manufacture bespoke devices.

Customer demand for the kind of problems that can be solved using software has always existed, but software solutions could only be profitably supplied as the cost of the necessary hardware continued to drop in price and increase in capacity (see Figure 1.1). The sequence of significant incremental improvements in hardware meant that customers found it cost effective to regularly replace existing hardware with something more recent. Figure 4.1 shows the growth in capacity of successive generations of memory devices; the memory capacity of individual devices increases over time.

Who is the customer and which of their demands can be most profitably supplied? These very tough questions are entrepreneurial and marketing problems and not the concern of this book.

Software ecosystems have been rapidly evolving for many decades and change is talked about as-if it were a defining characteristic, rather than a phase that ecosystems go through on the path to relative stasis. Change is so common that it has become blasé and on its own is no longer enough, now existing ways of doing things must be disrupted. The real purpose of disruption is redirection of all profits, from incumbents to those financing the development of disruptive systems. The fear of being disrupted by others is an incentive for incumbents to disrupt their own activities. Only the paranoid survive is a mantra for those striving to get ahead in a rapidly changing ecosystem.

The data processing and numerical calculation ecosystems<sup>552,632</sup> existed prior to the introduction of electronic computers; mechanical computers were used. Figure 4.2 shows UK government yearly expenditure on punched cards and mechanical data processing devices.

<sup>i</sup> The first hand-held computer was introduced around 1989 and vendors regularly introduced new products, claiming that customer demand now existed. Mobile phones filled a large enough demand that customers were willing to carry around an electronic device that required regular recharging; phones provided a platform for mobile computing that had a profitable technical solution.

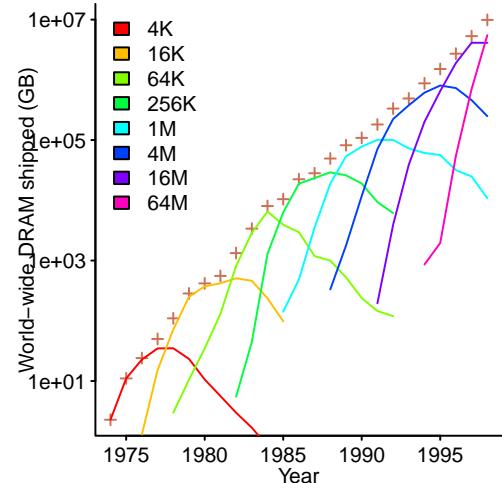


Figure 4.1: Total gigabytes of DRAM shipped world-wide in given year, along with shipments by device capacity (in bits). Data from Victor et al.<sup>1406</sup> code

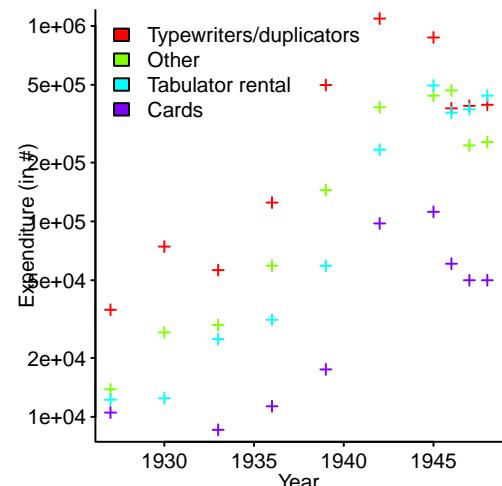
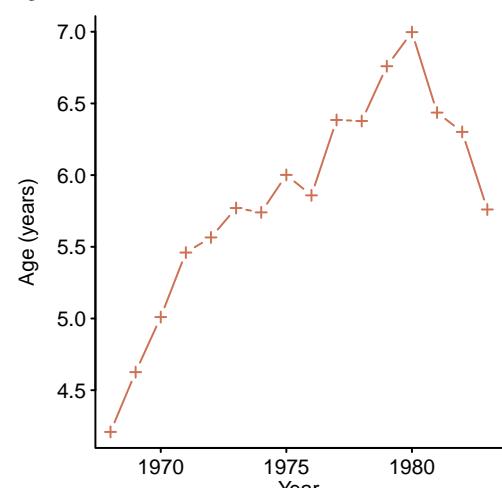


Figure 4.2: Yearly expenditure on punched cards and tabulating equipment by the UK government. Data from Agar.<sup>8</sup> code



The first release of a commercial software project implements the clients' understanding of the world, from some previous point in time. In a changing world, unimportant software systems have to adapt if they are to continue to be used; important software systems force the world to be adapted to use them.

Platform changes, hardware or software, provide opportunities for application changes...

Figure 4.3 shows the mean age of installed mainframes... This market is dominated by a single vendor (IBM), who decides product pricing and lifetime, it is not a market where competition provides customers a wider selection of options, e.g., fewer options for change are available... The increase in lifetime may be vendor driven, not customer driven...

The analysis in this book is oriented towards those involved in software development, rather than their customers, however it is difficult to completely ignore the entity that supplies the energy (e.g., money) needed for software ecosystems to continue to exist. The three primary ecosystems are the customers', the vendors' (i.e., companies engaged in economic activities, such as selling applications), and the developers' (i.e., people having careers in software development).

Figure 4.4 illustrates how one company, IBM, dominated the first 20+ years of the computer industry.

People who spend most of their time working within one ecosystem can have a blinkered view of the world in which they operate. An appreciation of the customer computer ecosystem, which may be very different from the ecosystem in which software is created, will help developers better tune what they do to the market in which it operates.

Connections between entities in an ecosystem are not always random. When network connections are driven by a non-random process, surprising properties can emerge, e.g., the *friends paradox*, where your friends have more friends, on average, than you do<sup>419</sup> and the *majority illusion*, where views about a complete network are inferred from the views of local connections.<sup>836</sup> In general, if there is a positive correlation, for some characteristic, between connected nodes, a node's characteristic will be less than the average of the characteristic for the nodes connected to it.<sup>390</sup>

The semiconductor industry has created an organization<sup>1206</sup> whose role is to publish agreed technological roadmaps outlining the characteristics of devices expected to be available in the coming years. The purpose of these roadmaps is to provide planning information, based on improvements that semiconductor manufacturers are working towards.

### 4.1.1 Evolution

Ecosystems contain a collection of interacting entities whose characteristics may change over time, including ceasing to exist (i.e., having a finite lifetime); ecosystems evolve over time and if the connections between geographical locations are weak, the paths followed in different locations can be different (e.g., evolution of religious beliefs<sup>1466</sup>).

The time changing behavior of ecosystems requires that they be modeled over time, a model built from measurements of a single snapshot in time can create a misleading picture because of survivorship bias. For instance, a snapshot of software systems currently being maintained is likely to find that the majority (51%, assuming a financial interest rate of 5%) have a ratio of total maintenance costs to total development costs greater than five (the ratio is greater than one for 80% of systems, greater than two for 68%; see [ecosystems/maint-dev-ratio.R](#)); however, the average ratio is 0.81.

Exponential improvement in product performance is not new. Figure 4.6 shows the increase in the maximum speed of human vehicles on the surface of the Earth and in the air over time.

Ecosystem evolution is path dependent.<sup>1302</sup> Things are the way they are today because past decisions selected particular paths to follow for particular reasons, e.g., the QWERTY keyboard layout was designed to reduce jamming in the early mechanical typewriters,<sup>314</sup> not optimizing typist performance (which requires a different layout).

Software ecosystems evolve when existing software is modified or when new software is written. The driver of change originates in customer demand, as perceived by those with the capacity to drive the changes (e.g., ability to make changes and/or funding to pay for them).

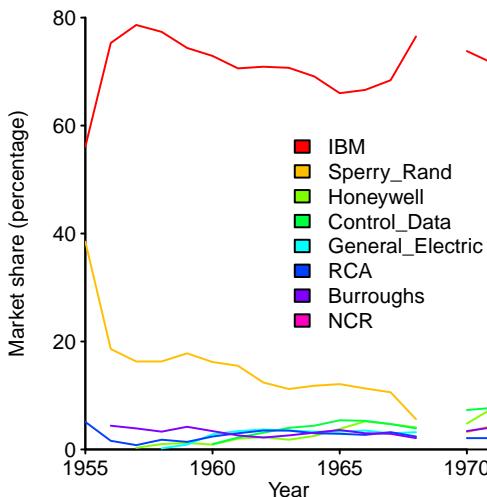


Figure 4.4: Computer installation market share of IBM and its top seven competitors (known at the time as the seven dwarfs; no data is available for 1969). Data from Brock.<sup>184</sup> [code](#)

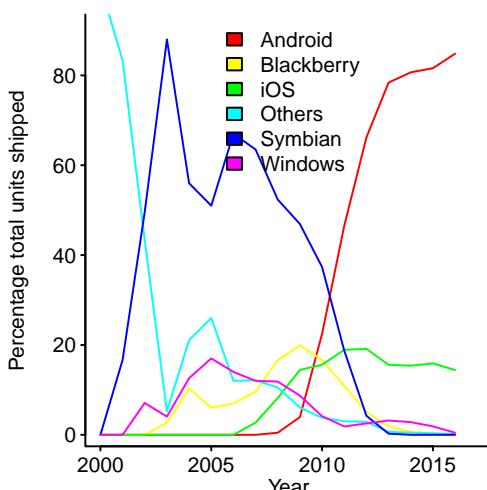


Figure 4.5: Mobile phone operating system shipments, as percentage of total per year. Data from Reimer<sup>686</sup> (before 2007) and Gartner<sup>480</sup> (after 2006). [code](#)

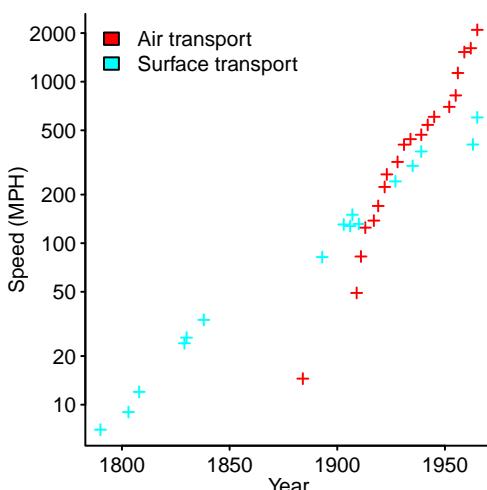


Figure 4.6: Maximum speed achieved by vehicles over the surface of the Earth and in the air, over time. Data from Lienhard.<sup>847</sup> [code](#)

Left untouched, software remains unchanged from the day it starts life, use does not cause it to wear out or break. However, the world in which software operates does not remain unchanged and it is this changing world that eventually breaks unchanged software. Software systems that have had little or no adaptation to a substantially changed world are sometimes known as *legacy systems*.

The incentives for investing in adapting existing software to a changed world include:

- continuing to make money through updates or sales of an existing product. Reasons for include being pushed by competition from other products keeping the Red Queen treadmill turning, and pushing customers to spend on upgrades (which may not fill any significant new customer need, but some markets have been trained, over decades of product releases, to believe that the latest version is always better than previous versions),
- if a potential new market opens up, a modified form of an existing product may be the most cost effective way of entering this market, e.g., the creation of a new processor creates an opportunity for a compiler vendor to add support for a new target instruction set,
- the cost of adapting existing software is less than the cost of not changing it, i.e., the job the software did, before the world changed, still has to be done,
- changes in the customer's world cause them to want functionality added to the software they current use...
- software developers feeling a need to change existing code (the desire to do something interesting is a significant motivation for some of those involved in the production of software systems),

A study by Keil, Bennett, Bourgeois, Garcá-Peña, MacDonald, Meyer, Ramirez and Yguel<sup>740</sup> applied various analysis techniques from ecology to Linux distributions. Figure 4.7 shows a phylogenetic tree of 56 Debian derived distributions, based on the presence/absence of 50,708 possible packages in each distribution.

The incentives for investing in creating new software include:

- wanting to enter a market and not having existing software that can be adapted to meet customer needs.
- the desire to create something new. This is a common human desire and in the case of researchers in academia is a requirement for anybody wanting to acquire or enhance an academic reputation.<sup>644</sup> Open source developers...

In software ecosystems, almost every major component has been rapidly evolving since computing began, e.g., hardware capability, vendors and people. Clients paying for software systems would like a quiet life, but have to stay on the Red Queen treadmill for fear that competitors will take away their customers. In the turmoil generated by change, people may have more opportunity for change than if they worked in mature industries.

Rapid hardware evolution appears to be coming to an end: increases in processor performance on an 18-month basis is now computing history. Figure 4.8 shows while the number of transistors in a device has continued to increase, clock frequency has plateaued (the thermal energy generated by running at a higher frequency cannot be extracted fast enough to prevent devices destroying themselves).

Software evolution is driven by a mixture of business and technical factors. A study by Branco, Xiong, Czarnecki, Küster and Völzer<sup>171</sup> analysed over 1,000 change requests in 70 business process models of the Bank of Northeast of Brazil. Figure 4.9 shows the distribution of the 388 maintenance requests made during the first three years of the Customer Registration project. Over longer time-scales the rate of evolution of some systems has a cyclic pattern.<sup>1487</sup>

Commercial viability of products that had been widely used...???

emailed and promised...??

What is the distribution of lifetimes of communities that have grown up around a particular software system?

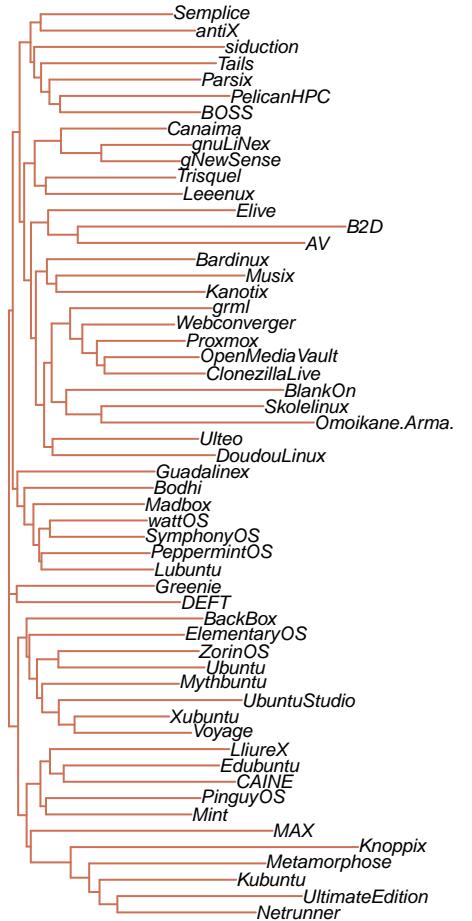


Figure 4.7: Phylogenetic tree of Debian derived distributions, based on 50,708 possible packages included in each distribution. Data from Keil et al.<sup>740</sup> [code](#)

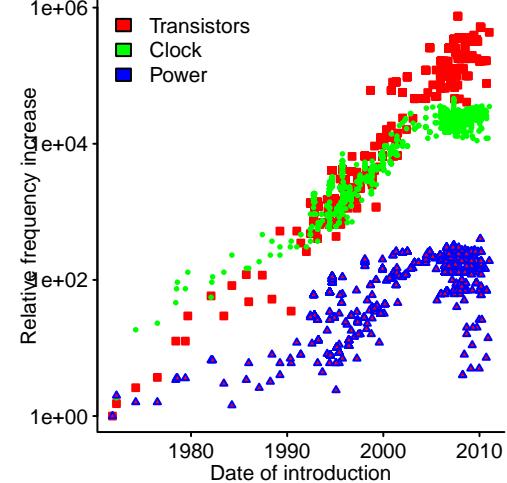


Figure 4.8: Number of transistors, frequency and SPEC performance of cpus when first launched. Data from Danowitz et al.<sup>312</sup> [code](#)

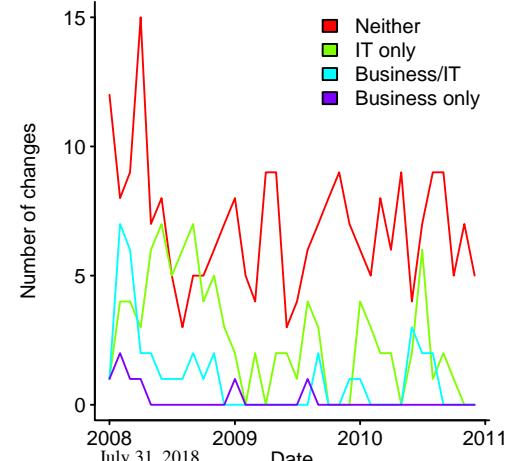


Figure 4.9: Number of process model change requests made in three years of a banking Customer registration

A study by Dunbar and Sosis<sup>364</sup> investigated human community sizes and lifetime. Figure 4.10 shows the number of founding members of 53 19th century secular and religious utopian communities, along with the number of year they continued to exist... .?

??,??

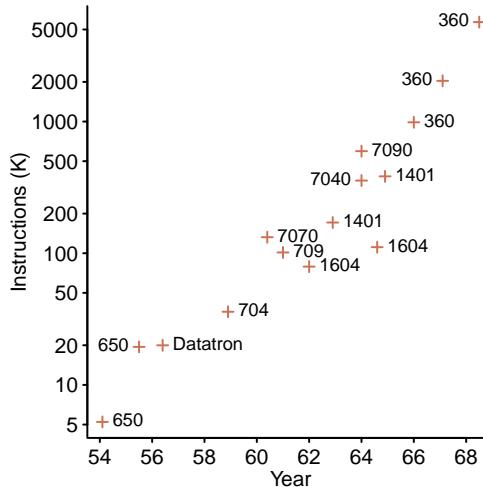


Figure 4.11: Total instructions in the software shipped with various models of IBM computer, plus Datatron from Burroughs. Data extracted from Naur et al.<sup>995</sup> code

## 4.1.2 Software ecosystems

Computers were originally delivered to customers as bare-metal (many early computers were designed and built by established electronics companies<sup>474</sup>); every customer wrote their own software, sometimes obtaining code from other users.<sup>ii</sup> As experience and customers accumulated,<sup>76</sup> vendors learned about the kinds of functionality that was useful across their customer base.<sup>199</sup> Figure 4.11 shows how the amount of code shipped with IBM computers increased over time; customer demand support interfaces to many different peripherals is a major driver of OS growth, see Figure 4.12.

An ecosystem contains entities of various sizes and if money can be made, then many companies are happy to be minor players. Minor players thrive by filling niches that are not worth the time and effort of the major players, or by having the agility to take advantage of short-term opportunities...

Having invested to create an ecosystem, vendors want to make it difficult for customers to switch to a different vendor. The availability of functionality that is unique to one ecosystem can both make it more attractive to customers and reduce customer choice (by increasing switching costs). Operating systems were the first software ecosystem and OS vendors offer unique functionality to third-party developers in the hope this will be used in the code they write. Developers may use the unique functionality because it is useful or through ignorance (i.e., they are not aware of the non-unique alternatives)...

Vendors wanting to sell their products on multiple operating systems have to decide whether to offer the same functionality across all versions of their product (i.e., not using functionality unique to one operating system to support functionality only available on that OS) or to provide some functionality that varies between operating systems.

The term *middleware* is sometimes applied to software designed to make it easier to port applications across different operating systems; the Java language is perhaps the most well-known example of middleware.

Operating system vendors dislike middleware because it makes it reduces customers' cost of switching vendors. Microsoft, with successive versions of Microsoft Windows, was the dominant OS vendor of the 1990s and 2000s and had an ecosystem control mechanism that was known as *embrace and extend*. Microsoft licensed Java and added Windows specific functionality to its implementation, which then failed to pass the Java conformance test suite. Sun Microsystems (who owned Java at the time) took Microsoft to court and won,<sup>1446</sup> Microsoft then refused to ship a non-extended version of Java as part of Windows, Sun filed an antitrust case and won.<sup>971</sup>

Figure 4.13 shows how the economic value, to software vendors, shifted from primarily bespoke development to package sales.

One technique of identifying economically significant ecosystems is monitoring job adverts specifying specific skills...

Total US expenditure on software 1959—1998...?.

### 4.1.2.1 Countries

Governments are aware of the importance of having national software ecosystems,<sup>1029</sup> both in economic terms (e.g., industry investment in software systems<sup>246</sup> that keep them competitive) and as a means of self-determination (i.e., not having important infrastructure dependent on companies based in other countries); there is no shortage of recommendations<sup>1323</sup> for how to nurture IT-based business.

The software export figures given for a country can be skewed by a range of factors and the headline figures may not include associated costs.<sup>596</sup>

<sup>ii</sup> Software user-groups are almost as old as computers.<sup>50</sup>

### 4.1.3 Evolving data about evolving systems

Measurements of evolving ecosystems are also going to evolve, and speed of evolution may mean that patterns found in the data change equally quickly.

Figure 4.14 shows a sorted list of the total amount earned by individuals through bug bounties programs. Both studies downloaded data that was available on the HackerOne website; the study by Zhao, Grossklags and Liu<sup>1490</sup> used data from November 2013 to August 2015, and the study by Maillart, Zhao, Grossklags and Chuang<sup>888</sup> used data from March 2014 to February 2016.

Applying the results from any analysis of an evolving system needs to take into account any evolution that has occurred since the measurements used in the analysis were gathered...

## 4.2 Culture

Culture<sup>iii</sup> improves adaptability. Culture is common in animals, but cultural evolution<sup>938</sup> is rare; perhaps limited to humans, song birds and chimpanzees.<sup>165</sup>

People and existing code are the carriers of software culture.

Learning by observing others, *social learning*, enables animals to avoid the potentially high cost of *individual learning*.<sup>47</sup> Average population fitness is increase when members are capable of deciding which of two learning strategies, social learning or individual learning, is likely to be the most cost effective<sup>166</sup> (i.e., the cost of individual learning can be focused where the benefit is likely to be maximised).

Milk used to be left in open bottles on customer doorsteps, in England; various species of birds were known to drink some of this milk. When the practice changed, to sealing bottles (these days with aluminium foil), some birds learned to peck open milk bottle tops, with the blue tit being the primary scavenger.<sup>438iv</sup> The evidence points towards those bird species that forage in flocks have the ability to learn both socially and individually, while species that are territorial (i.e., chase away other members of their species) primarily use individual learning.<sup>829</sup>

Social learning is a skill in which 2.5-year-old children significantly outperform adult our closest primate relatives, chimpanzees and orangutans<sup>608</sup> (performance on other cognitive tasks was broadly comparable). emailed for data...

Just using social learning is only a viable strategy in an environment that is stable between generations of learners; if the environment is likely to change between generations, then copying runs the risk of learning skills that are no longer effective. Analysis of a basic model<sup>918</sup> shows that for a purely social learning strategy to spread in a population of individual learners, the probability of environmental change in any given generation,  $u$ , and learning costs,  $b$  is the cost of individual learning and  $c$  the reduction in the cost of learning achieved by social learning, the following relation must be true:  $u > \frac{c}{b}$ .

Discoveries and inventions are often made by individuals and it might be expected that larger populations will contain more tools and artefacts and have a more complex cultural repertoire than smaller populations.<sup>833</sup> The evidence is mixed, with population size having a positive correlation with tool complexity in some cases;<sup>275</sup> it has been proposed<sup>602</sup> that the indigenous people's of Tasmania lost valuable skills and technologies because their effective population size shrunk when the sea level rose, cutting the islander's off from contact with mainland Australia.

Experimental studies have found conventions spontaneously emerging in networks, having various topologies, containing up to 48 members (the maximum experimental condition).<sup>225</sup> One advantage of agreed conventions is a reduction in effort required to communicated when performing a joint task (e.g., the number of words used<sup>262</sup>)

Useful new knowledge is not always uniformly diffused out into the world, to be used by others; a constantly changing environment introduces uncertainty<sup>1119</sup> (i.e., noise is added to

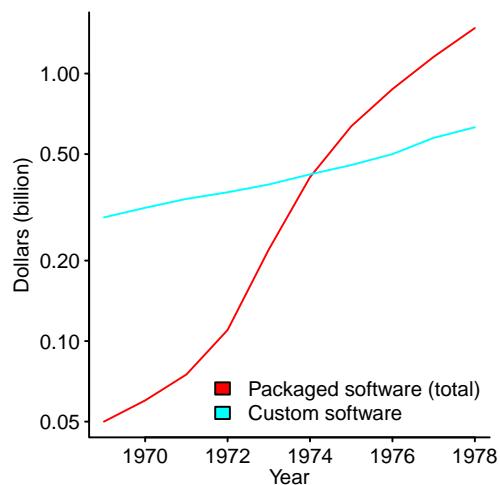


Figure 4.13: Total value of custom and packaged software (hardware vendor+third-party) sales in the US. Data from Phister.<sup>1087</sup> [code](#)

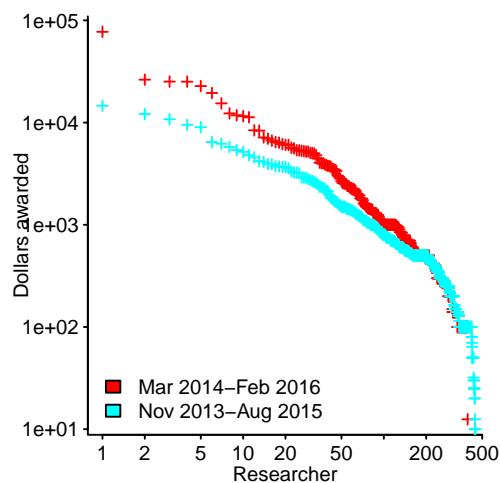


Figure 4.14: Sorted list of total amount awarded by bug bounties to individual researchers, based on two datasets downloaded from HackerOne. Data from Zhao et al<sup>1490</sup> and Maillart et al.<sup>888</sup> [code](#)

the knowledge signal) and influential figures may suppress use of the ideas (e.g., the spread of Feynman diagrams<sup>868</sup>).

Conformist transmission is the name given to the hypothesis that individuals possess a propensity to preferentially adopt the cultural traits that are most frequent in the population. Under conformist transmission, the frequency of a trait among the individuals within the population provides information about the trait's adaptiveness. This psychological bias makes individuals more likely to adopt the more common traits than they would under unbiased cultural transmission. Unbiased transmission may be conceptualized in several ways. For example, if an individual copies a randomly selected individual from the population, then the transmission is unbiased. If individuals copy their parents or just their mother, then transmission also is unbiased.

At the population level, conformist transmission causes more common traits to increase in frequency. If cultural transmission is unbiased, then, barring the action of other forces, transmission will leave the frequency of the traits unchanged from one generation to the next. For example, if 60% of a population is performing a certain behavior, barring other forces, 60% of the population in the next generation also will perform that behavior. In contrast, conformist transmission would increase the frequency of the trait from 60% in one generation to, say, 65% in the next generation. All other factors being equal, the frequency of the most prevalent trait will continually increase from one generation to the next. If it were the only transmission bias, conformist transmission would rapidly cause the most frequent cultural traits to become the only cultural traits.

While overconfidence at the individual level decreases the fitness of the entrepreneur (who does not follow the herd, but tries something new), the presence of entrepreneurs in a group increases group level fitness (by providing information about the performance of new ideas).<sup>133</sup>  
emailed...???

?

Size of crowd looking up increases with size of stimulus crowd<sup>945</sup> ...

Fashion as a prestige signalling system... Following leaders generates a pressure to conform...?

Hype cycles...??

Self-enforcing norms<sup>226</sup>...

Giving more attention to high-status individuals... A study by Simcoe and Waguespack<sup>1255</sup> ...

### 4.2.1 Software culture

Coding was originally classified as a clerical activity (a significant under-appreciation of the cognitive abilities required) and the sexist division of working roles prevalent during the invention of electronic computing specified coding as a job for women; many of those working as programmers on the early computers were women (mens' work included specifying the formulas to be used...)<sup>849</sup>

Is there a software culture shared across many ecosystems, or does the culture of every major development/customer ecosystem evolve completely independently? Perhaps software system development has changed so rapidly that cultural behaviors have not had time to spread and become widely adopted, before others take their place. Distinct development practices did evolve at the three early computer development sites in the UK.<sup>214</sup>

Software cultures will include aspects of the application domain culture in which the software is produced... Character set encodings are a basic component of software culture. These have been evolving since the 1870s,<sup>436</sup> with the introduction of the telegraph; early computer capacity restrictions drove further evolution<sup>882</sup> and the lifting of these restrictions allowed the spread of a single encoding for all the World's characters.<sup>1378</sup>

The culture of the society in which developers grew up and were educated will have instilled specific non-software usage habits. Metaphors<sup>803</sup> are a figure of speech which apply the features associated with one concept to another concept. For instance, concepts involving time are often expressed by native English speakers using a spatial metaphor. These metaphors take one of two forms—one in which time is stationary, and we move through it (e.g., "we're

approaching the end of the year"); in the other case, we are stationary and time moves toward us (e.g., "the time for action has arrived").

A study by Boroditsky<sup>157</sup> investigated subject's selection of either the ego-moving or the time-moving frame of reference. Subjects first answered a questionnaire dealing with symmetrical objects moving to the left or to the right; the questions were intended to prime either an ego-moving or object-moving perspective. Subjects then read an ambiguous temporal sentence (e.g., "Next Wednesday's meeting has been moved forward two days") and were asked give the new meeting day. The results found that 71% subjects responded in a prime-consistent manner; of the subjects primed with the ego-moving frame, 73% thought the meeting was on Friday and 27% thought it was on Monday. Subjects primed with the object-moving frame showed the reverse bias (i.e., 31% and 69%).

Native Chinese speakers also use spatial metaphors to express time related concepts, but use the vertical axis rather than the horizontal axis used by native English speakers.

Cultural conventions can be domain specific; for instance, in the US politicians *run* for office, while in Spain and France they *walk*, and in Britain they *stand* for office. These metaphors may crop up as supposedly meaningful variable names, e.g., `ran_for`, `is_standing`.

Is English the lingua-franca of software developers?

?

The office suite LibreOffice was originally written by a German company, and many comments were written in German. A decision was made that all comments should be written in English; Figure 4.15 shows the decline in the number of comments written in German.

Chinese developers required to speak English at work...<sup>1424</sup> emailed for data...

In ALGEC,<sup>634</sup> a language invented in the Soviet Union, keywords can be written in a form that denotes their gender and number. For instance, Boolean can be written: ?????????? (neuter), ?????????? (masculine), ?????????? (feminine) and ?????????? (plural). The keyword for the go to token is `to`; something about Russian makes use of the word go unnecessary.

A fixation on a particular way of viewing the world is not limited to language conventions, it can also be found in mathematics. For instance, WEIRD people have been drilled to use a particular algorithm for dividing two numbers and believe that the Romans had serious difficulties with division because of the system of number representation they used. Division of Roman numerals is simple if the appropriate algorithm is used:<sup>7</sup>...

When asked to name an object or action, people have been found to give a wide range of different names. A study by Furnas, Landauer, Gomez, and Dumais<sup>465,466</sup> described operations to subjects who were not domain experts (e.g., hypothetical text editing commands, categories in *Swap 'n Sale* classified ads, keywords for recipes) and asked them to suggest a name for each operation. The results showed that the name selected by one subject was, on average, different from the name selected by 80% to 90% of the other subjects (one experiment included subjects who were domain experts and the results for those subjects were consistent with this performance). The frequency of occurrence of the names chosen tended to follow an inverse power law.

Object naming has been found to be influenced by recent experience,<sup>1270</sup> practical skills (e.g., typists selecting pairs of letters that they type using different fingers<sup>1385</sup>) and egotism (e.g., a preference for letters in one's own name or birthday related numbers<sup>763,1024</sup>).

Separating behavior that is the result of cultural learning rather than developers finding a common solution can be difficult. For instance, a study of the use of single letter identifiers in five languages<sup>125</sup> found that `i` was by far the most common in programs written in four of the languages. Is this usage primarily driven by developers abbreviating the words `integer` (the most common variable type) or `index`, or by seeing this usage in example code in books and on the web?...

## 4.2.2 Folklore

The dearth of experimental evidence has left a vacuum that has been filled by folklore and old-wives' tales. Examples of software folklore include claims of a 28-to-1 productivity difference between best/worst developers and that parameter passing in function calls is resource intensive in embedded systems.

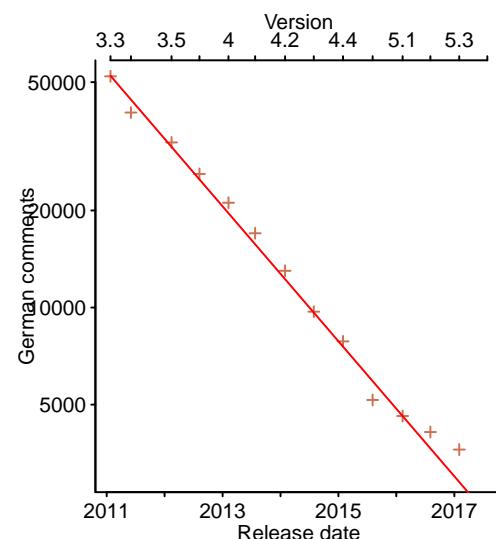


Figure 4.15: Estimated number of comments written in German, in the LibreOffice source code. Data from Meeks.<sup>940</sup> [code](#)

The productivity claim, sometimes known as the *Grant and Sackman study*, is based on an incorrect reading of a 1966 study<sup>538</sup> by these two authors. This study set out to measure the extent to which the then new time-sharing approach to computer usage was more productive for software development than existing batch processing systems (where jobs were typically submitted via punched cards, with program executing and their output printed on paper several hours later).

This paper<sup>538</sup> was published in a widely read journal and summarised the data; a table listed the ratio between the best subject performance using the fastest system and the worst subject performance on the slowest system; the 28:1 ratio included programmer and systems performance differences (if batch/time-sharing differences are separated out the maximum difference ratio is 14:1, the minimum 6:1). The actual measurement data was published in a relatively inaccessible journal.<sup>1190</sup>

In 1981, a study by Dickey<sup>350</sup> separated out subject performance from other factors, adjusted for individual subject differences and found a performance difference ratio of 1:5. However, by 1982 the 28:1 ratio had appeared in widely read magazine articles and books and had become established as *fact*. In 1999, a study by Prechelt<sup>1114</sup> explained what had happened, for a new audience (perhaps if Figure 8.21 had appeared in the second 1966 paper...).

Embedded software runs on resource limited hardware, which is often mass-produced and saving pennies per device can add up to lots of money; systems are populated with the smallest possible memory and power consumption is reduced by using the slowest possible clock speeds, e.g., closer to 1 MHz than 1 GHz.

Experienced embedded developers are aware of the hardware performance limitations they have to work within. Many low cost processor have a very simple architecture with relatively few instructions and parameter passing, to a function, can be very expensive (in execution time and code size) compared to passing values to functions in global variables on some processors.

A study by Engblom<sup>388</sup> investigated differences in the characteristics of embedded C software and the SPECint95 benchmark. Figure 4.16 shows the percentage of function definitions containing a given number of parameters, for embedded software the SPECint95 benchmark and desktop software measured by Jones.<sup>699</sup> A Poisson distribution provides a reasonable fit to both sets of data; for desktop software, the distribution of function definitions having a given number of parameters the Poisson distribution has  $\lambda = 2$ , while for embedded developers  $\lambda = 0.8$ .

These measurements were of source code from the late 1990s, are things likely to have changed?

Companies are likely to be just as interested in saving pennies. Compilers may have become better at reducing function parameter overheads for some processor, however, it is beliefs that drives developer usage.

Embedded devices have become more mainstream, with companies selling IoT devices with USB interfaces. This availability provides an opportunity for aspects of the culture of developing for desktop and mobile systems to invade the culture of embedded development. In situations where code size or/and performance is critical, developers looking for savings may learn about the overheads of parameter passing.

Is the values of  $\lambda$  always approximately 0.8 or 2.0? Perhaps there is a range of values, depending on developer experience (old habits die hard and parameter overhead will depend on processor characteristics, e.g., 4-bit, 8-bit and 16-bit processors).

### 4.2.3 Expertise

What is expertise and how might people become experts? Expertise might be defined by what somebody can do, i.e., performance based, or by what other people think, i.e., social based:

- performance defined experts: somebody who can perform a task better than the average person on the street, better than most other people who can also perform the task, or in some chosen top percentage,
- socially defined experts: these include well-established figures, perhaps holding a senior position with an organization heavily involved in that domain, and self-proclaimed experts,

who are willing to accept money from clients who are not willing to take responsibility for proposing what needs to be done<sup>51</sup> (e.g., the role of court jester who has permission to say what other cannot...).

There are domains in which those acknowledged as experts do not perform significantly better than those considered to be non-experts<sup>212</sup> in some cases non-experts have been found to outperform experts within their domain.<sup>1452</sup> An expert's domain knowledge can act as a mental set that limits the search for a solution; the expert becomes fixated within the domain; in cases where a new task does not fit the pattern of highly proceduralized behaviors of an expert, a novice has an opportunity to do better.

Expertise within one domain does not confer any additional skills within another domain.<sup>?</sup>

Studies have found that the main factor in acquiring expert performance is time spent in *deliberate practice*.<sup>392</sup> Deliberate practice is different from simply performing the task, it requires that people monitor their practice with full concentration and receive feedback<sup>631</sup> on what they are doing (often from a professional teacher). The goal of deliberate practice being to improve performance, not to produce a finished product.

Studies of the backgrounds of recognized experts, in many fields, found that the elapsed time between them starting out and carrying out their best work was at least 10 years, often with several hours of deliberate practice every day of the year.<sup>393</sup>

Become a performance defined expert a person needs motivation, time, economic resources, an established body of knowledge to learn from, and teachers to guide; while learning, performance feedback is needed.

An established body of knowledge to learn from requires a domain to have existed in a sufficiently stable state for long enough for a proven body of knowledge to have been established. The availability of teachers requires the domain be sufficiently stable that most of what potential teachers have learned is still applicable to students; if the people with the knowledge and skills are to spend a substantial amount of time teaching, they need to be paid enough to make a living.

Software developers are not professional programmers any more than they are professional typists. Reading and writing software is one aspect of building software systems and developers have to compare the cost/benefits of spending time becoming more skilful at the various tasks required, including programming and the application domain in which their software is used.

Computer users have distinct command usage habits,<sup>1211</sup> but a repertoire of finely tuned habits does not make a person an expert...

While the history of software development is a few human generations old, there have been a steady stream of substantial changes; substantial change is written about as-if it is the norm. Anybody who invests in many years of deliberate practice on a specific technology may find there are few customers for the knowledge and skill acquired, i.e., are willing to pay a higher rate to do the job, than that paid to somebody with a lot less expertise...

What level of software expertise is worth acquiring in a rapidly changing ecosystem? The level of skill required to be employed in software development is relative to everybody who applies for the job... In an expanding market employers may have to make do with whatever they can get... Those with higher skill levels may have the luxury of being able chose the work that interests them...

When estimating benefit over a relatively short time frame, time spent learning more about the application domain frequently has a greater return than honing programming skills. and the data for this...

Estimating the effort needed to implement some software functionality or system is a common requirement. Is project estimation a skill worth investing in to acquire? Would a person who invests in acquiring such a skill improve their employability prospects or earn a larger salary?...

Based on the available evidence, experience writing software is not a reliable indicator of the accuracy of a person's effort estimation...

#### 4.2.4 Organizational learning and forgetting

The performance of organizations, like that of individuals (see Figure 2.17), improves with practice; organizations also forget (in the sense that performance on a previously learned task degrades with time). Industrial studies have focused on learning by doing,<sup>1354</sup> that is the passive learning that occurs when the same, or very similar, product is produced over time.

The impact of organizational learning during the production of multiple, identical units (e.g., a missile or airplane) can be modeled to provide a means of estimating the likely cost and timescale of producing more of the same units.<sup>511</sup> Various models of organizational production<sup>647,924</sup> based on connected components, where people are able to find connections between nodes that they can change to improve production, produce the power laws of organizational learning found in practice. other models...

There are trade-offs to be made in investment in team learning; there can be improvements in performance and term performance can be compromised.<sup>201</sup>

A study by Nembhard and Osothsilp<sup>1000</sup> investigated the impact of learning and forgetting on the production time of car radios; various combined learning-forgetting models have been proposed and the quality of fit to the data was compared. Figure 4.17 shows the time taken to build a car radio against cumulative production, with an exponential curve fitted to each period of production (break duration above the x-axis). Note, build time increases after a break and both a power law and exponential have been proposed as models of the forgetting process.

Software is easily duplicated, continual reimplementation only occurs in experiments (see Figure 2.18). Software systems also tend to be more complicated than radios and take longer to implement.

Figure 4.18 shows the man-hours needed to build three kinds of ships, 188 in total, at the Delta Shipbuilding yard, between January 1942 and September 1945. As experience is gained building Liberty ships the man-hours required, per ship, decreases.<sup>1354</sup> Between May 1943 and February the yard had a contract to build Tankers, followed by a new contract to build Liberty ships and then Colliers. When work resumes on Liberty ships, the man-hours per ship is higher than at the end of the first contract, presumably some organizational knowledge about how efficiently build this kind of ship had been lost.

What opportunities are there for organizational learning to occur during the production of software systems?... 7Digital data?...

### 4.3 Customer ecosystems

Software may be written for a single customer or with the intent of selling copies to multiple customers. Single customers may be large organizations paying millions to solve a major problem, or a single developer out to enjoy himself.

The military were the first customers for computers and financed the production of one-off systems.<sup>440</sup> The first commercial computer, the Univac I, was introduced in 1951 and 46 were sold; the IBM 1401,<sup>479</sup> introduced in 1960, was the first computer to exceed one thousand installations, with an estimate 8,300 systems in use in July 1965.<sup>879</sup> During the 1950s a steady stream of new computers were introduced, with many customers initially preferring to rent (management feared equipment obsolescence in a rapidly changing market and had little desire to be responsible for maintenance<sup>798</sup>) and vendors nearly always preferring to rent (a license agreement could restrict customers' ability to connect cheaper peripherals to equipment of they do not own) preferred to rent. Figure 4.19 shows the shift from rental to purchase of computers by the US Federal Government.

The U.S. Government was the largest customer for computing equipment during the 1950s and 60s, and enacted laws to require vendors to supply equipment that conformed to various specified standards, e.g., the Brooks Act<sup>1358</sup> in 1965. The intent of these standards was to reduce costs, to the government, by limiting vendors ability to make use of proprietary interfaces to restrict competition.

The customers for these expensive computers were large organizations who had to regularly perform simple numeric calculations on large scale, e.g., payroll (large companies employ

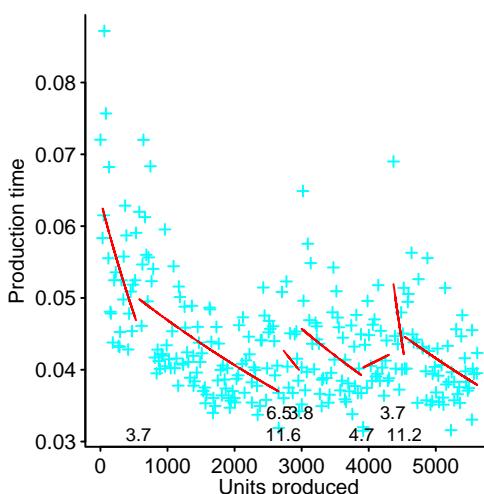


Figure 4.17: Hours required to build a car radio after the production of a given number of radios, with break periods (shown in days above x-axis); lines are models fitted to each production period. Data extracted from Nembhard et al.<sup>1000</sup> code

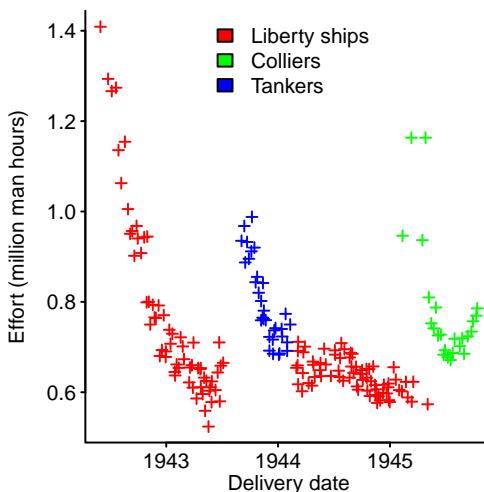


Figure 4.18: Man-hours required to build a particular kind of ship, at the Delta Shipbuilding yard, delivered on a given date (x-axis). Data from Thompson.<sup>1354</sup> code

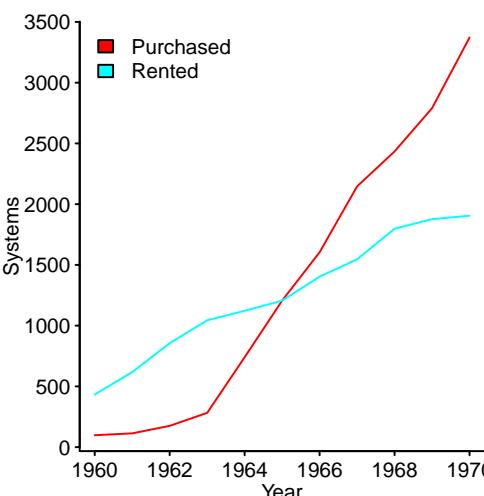


Figure 4.19: Total computer systems purchased and rented by the US Federal Government in the respective fiscal years ending June 30. Data from US Government General Accounting Office.<sup>280</sup> code

thousands of people, each requiring a unique weekly or monthly payroll calculation<sup>546</sup>), making forecasts and calculating engineering tables.

Large organizations had problems that were large enough to warrant the high cost of buying and operating a computer.<sup>1233</sup> A computer services industry<sup>1475</sup> quickly grew (during 1957, in the US, 32 of the 71 systems offered were from IBM<sup>878</sup>), providing access to computers in central locations (often accessed via dial-in phone lines); with processing time rented by the hour.<sup>62</sup> Computing as a utility service for general use, like electricity, looked to some like a viable business model.<sup>1226</sup> Figure 4.20 shows, for the US, service business revenue in comparison to revenue from system sales.

The introduction of microcomputers decimated the computer services business;<sup>215</sup> the cost of buying a microcomputer was could be less than the monthly rental to a service bureau.

What problems are customers interesting in solving? Survey of applications used<sup>798</sup>...

The IBM 360, initially delivered in 1965, was the first compatible computer family, that is successive generations could run software that ran on earlier machines...

How much money do different customer ecosystems spend on software? Estimating UK investment in intangible assets and Intellectual Property Rights<sup>517</sup>... Economically significant customer ecosystems include:

- desktop computing:...
- embedded systems:... many military projects contain a substantial embedded component, Figure 4.24...
- games consoles: see Figure 3.12...
- office computing:... government... serving the needs of running a country,
- scientific/engineering computing: the substantial number of calculations required to obtain answers for some economically important problems has created a niche market for compute intensive systems, known as *super computers*.

Initially powerful single processors were used, running for long periods of time, some problems (e.g., weather prediction) can be broken down into subcomponents which makes it possible to use multiprocessor systems executing in parallel. use of distributed computing resources, e.g., systems designed to solve a single problem, e.g., ANTON<sup>1235</sup> is built from ASICs to solve problems in molecular dynamics, repurposing of GPUs... Rlang-/Top500.R.

- Bespoke systems: software systems produced to order by one customer who pays for everything. Almost as soon as computers became available million line programs were being written to order. Figure 4.21 shows lines of code and development costs for US Air Force software projects, by year, excluding classified projects and embedded systems; the spiky nature of the data suggests that LOC and development costs are counted in the year a project is delivered.

In-house software... UK government data...<sup>140</sup>

?

emailed for data...?

Customers prefer to use generally available hardware and software. Competition helps stop prices spiralling upwards, general availability increases the likelihood that new staff will already be familiar, easier to justify purchase choice of a well-known name. Disadvantage of general availability is that the market may move in a direction not best suited...

Some applications run directly on the underlying hardware, others rely on an operating system to manage basic housekeeping activities, while others execute in an ecosystem containing thousands of support programs that are bundled with an operating system.

Users running applications may be unaware of the underlying operating system, but from the developers' perspective the operating system in an important defining characteristic of a customer's software platform.

Workload analysis,<sup>417</sup> analyzing computer usage and balancing competing requirements to maximise throughput, has always been an important job in any large computer installation...  
TODO? [ecosystems/CSD-95-887.R](#)

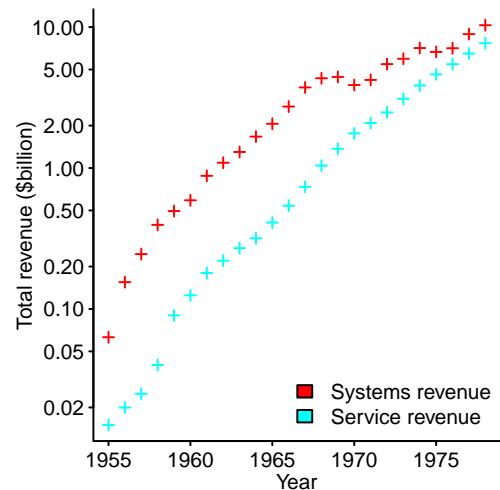


Figure 4.20: Total U.S. revenue from sale of computer systems and data processing service industry revenue. Data from [DOE](#) Pinter<sup>1087</sup>, table II.1.20 and II.1.26. [code](#)

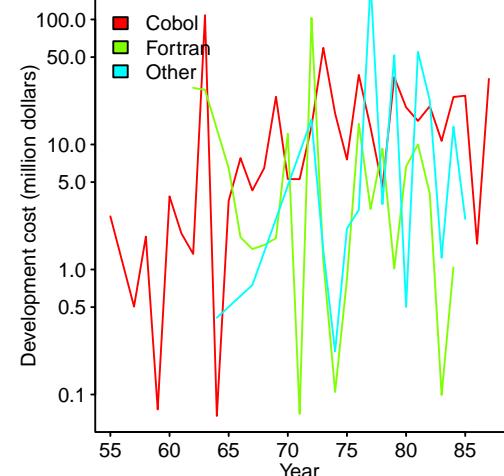
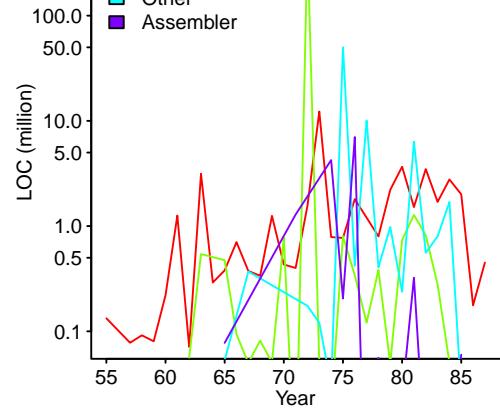


Figure 4.21: Yearly development cost and lines of code delivered to the US Air Force between 1960 and 1986. Data extracted from NeSmith<sup>1001</sup>. [code](#)

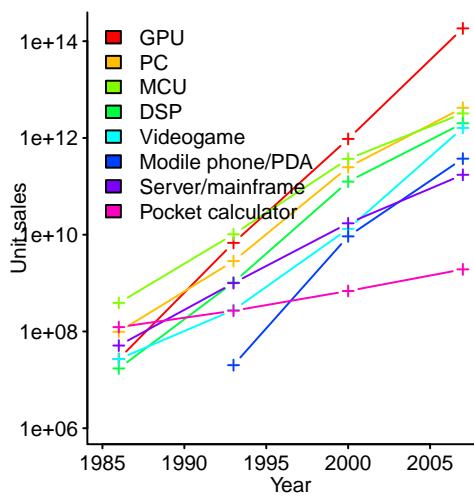


Figure 4.22: Total sales of various kinds of processors. Data from Hilbert et al.<sup>616</sup> [code](#)

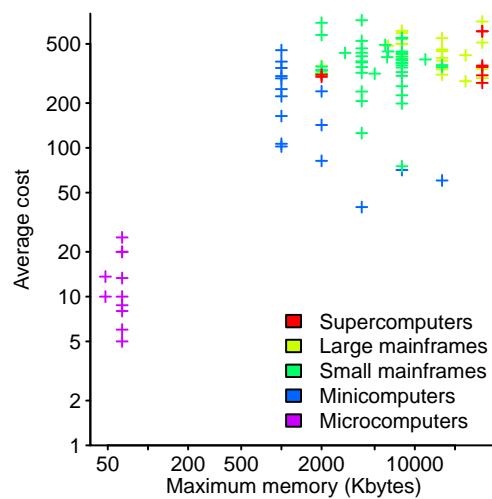


Figure 4.23: Maximum memory capacity and average cost of computer systems in 1981. Data from Ein-Dor.<sup>379</sup> [code](#)

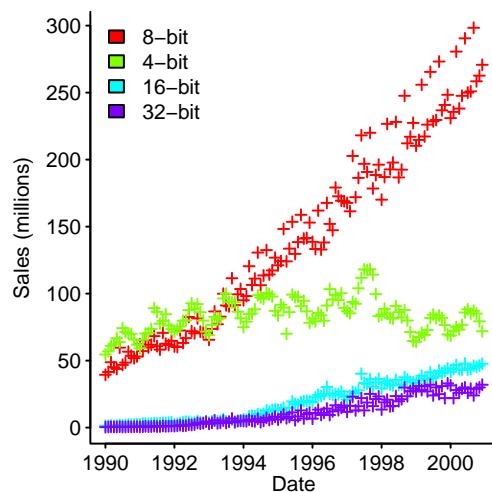


Figure 4.24: Monthly unit sales (in millions) of microprocessors having a given bus width. Data kindly supplied by Turley.<sup>1369</sup> [code](#)

### 4.3.1 Hardware ecosystems

Hardware is needed to execute software<sup>v</sup> and hardware characteristics and limitations can have important consequences for the design of systems.

The classification of computer hardware into mainframes,<sup>vi</sup> minicomputers<sup>7</sup> and microcomputers<sup>813</sup> is primarily marketing driven<sup>vi</sup>, with each platform class occupying successively lower price points and targeting different kinds of customers (e.g., mainframes large businesses, minicomputers technical and engineering companies, and micros individuals). Two hardware ecosystems that receive less publicity are supercomputers<sup>120</sup> (i.e., the fastest computers of the day, often used for scientific and engineering calculations) and embedded systems (where the computing aspect is often invisible ... of which mobile and IoT...; see Figure 4.22 and Figure 4.24)...

For many decades large manufacturers of computing systems periodically introduced new product ranges containing cpus<sup>108</sup> that were incompatible with previous product ranges. Even IBM, known for the compatibility of its 360/370 family of computers, continued to regularly introduce new systems based on incompatible cpus.

Software was restricted by both compute performance and available memory... Figure 4.23 shows the maximum memory capacity of systems available in 1981 and their average cost (which includes peripherals)...

? [ecosystems/CompWorld85.R](#)<sup>601</sup>

For its x86 family of processors, Intel made upwards compatibility a requirement<sup>vii</sup>. An apparently insatiable market demand for faster processors and large sales volume created the incentive to continually make significant investments in building faster processors; see Figure 4.22. The x86 family steadily increased market share,<sup>1305</sup> to eventually become dominant in the non-embedded computing market; one-by-one manufacturers of other processors ceased trading and the cpus they sold were donated to museums.

The market dominance of IBM hardware and associated software (50 to 70% of this market during 1969-1985;<sup>362</sup> see Figure 1.5) is something that developers now learn through reading about the history of computing. However, the antitrust cases against IBM continue to influence how regulators think about how to deal with monopolies in the computer industry and on how large companies structure their businesses.<sup>1062</sup>

While the practices and techniques used during one hardware era (e.g., mainframes, minicomputers, microcomputers, the internet) might not be appropriate in a later era, they have left their mark on software culture in the form of language standards written to handle more diverse hardware than exists today<sup>699</sup>... Also, each hardware generation has tended to be initially populated by developers who are relatively new to computing and thus unfamiliar with what has gone before and so often reinvented things...

Changes in the relative performance of various components impacts how systems are optimally designed and built, which in turn impacts software design choices which may take many years to catch up. data on sort algorithms that are cache dependent...

Embedded systems is a software ecosystem that has not attracted as much attention from the academic software engineering research community as desktop systems. Figure 4.24 shows that in terms of microprocessor sales volume the embedded systems market is significantly larger than what might be called the desktop computer market. mobile market...

The total energy consumed by computing devices is increasing in importance to customers, from laptop and mobile phone battery usage,<sup>91</sup> compute infrastructure within a building<sup>738</sup> to the design of Warehouse-scale systems<sup>97</sup> (there are large energy efficiency gains to be had running software in the Cloud<sup>902</sup>). The world-wide energy consumed by computing systems is...

The larger the computing system, the more power it consumes [benchmark/brl/brl.R](#).

The issue of significant variation in power consumption across different chips from the same product, when executing the same program, is discussed elsewhere...

<sup>v</sup> The division between hardware and software can be very fuzzy; for instance, the hardware for Intel's Software Guard Extensions (SGX) instructions consists of software micro operations performed by lower level hardware.<sup>294</sup>

<sup>vi</sup> The general characteristics of the central processors and subsystems are remarkable similar, and they follow similar evolutionary paths because they are solving the same technical problems.

<sup>vii</sup> To the extent of continuing to duplicate faults in earlier processors.<sup>7</sup>

Figure 4.25 shows how wafer production revenue at TSMC (the world's largest independent semiconductor foundry) has migrated to smaller process technologies over time and how demand has shifted across major market segments...

## 4.4 Vendor ecosystems

### 4.4.1 Companies

A company is a legal entity that limits the liability of the directors and owners; the directors of a company are required to maximise the return on investment to shareholders... Commercial pressure to deliver results in the short term rather than taking a longer term view ...<sup>42</sup>

The first software company, selling software consulting services, was founded in March 1955.<sup>787</sup> Commercial software products, from third-parties, had to wait until computer usage became sufficiently widespread that a profitable market for them was considered likely to exist. One of the first third-party software products ran on the RCA 501, in 1964, and created program flowcharts.<sup>692</sup>

The birth, growth and death of companies<sup>771</sup> is of economic interest to governments seeking to promote the economic well-being of their country. Company size, across countries and industries, roughly follows a Pareto distribution,<sup>326</sup> while company age has an exponential distribution;<sup>267</sup> most companies are small and die young. and for software companies?...

Figure 4.26 shows the number of UK companies registered each month having the word software or computer in their SIC description (age distribution coming...).

Markets for particular classes of products (e.g., electric shavers, jet engines and video cassette recorders) and services evolve, with firms entering and exiting. Manufacturing markets have been found to broadly go through five stages, from birth to maturity.<sup>9</sup>

The invention of the Personal Computer created a new ecosystem and people formed new companies to seek their fortune in the markets created. Figure ?? shows how the growth and consolidation of PC manufacturers operating in this ecosystem followed a similar pattern to what occurred after the invention of the automobile...

Top European software companies, by turnover Truffle100/...

The commercial software ecosystem contains companies providing services such as consulting, training and recruitment.<sup>268</sup> Companies who create software systems divide into those that create bespoke systems, paid for by a single customer, and those that create a product to sell to multiple customers...

The development of a software system can involve a whole ecosystem of supply companies. For instance, the winner of a contract to develop a large military system often subcontracts-out work for many of the subsystems to different suppliers; each supplier using different computing platforms, languages and development tools (Baily et al<sup>85</sup> describes one such collection of subsystems).

Companies may operate in specific customer ecosystems or within a software ecosystem of a large company might be defined by the platform on which they run, e.g., software running on Microsoft Windows, by the kind of work users of the software do, e.g., making presentations, or by the industry in which users of the software operate, e.g., the travel business.

A study by Crooymans, Pradhan and Jansen<sup>299</sup> investigated the relationship connections between companies in a local software business network; Figure 4.27 shows the relationship links between companies, with a few large companies and many smaller ones.<sup>viii</sup>

Software startups take many forms, from a pseudo independent group working within a large company to a few people starting a new company. Reasons for forming a startup vary from wanting to make a living running a business, to being funded by Venture Capital (VC) to the tune of many millions (Martínez<sup>901</sup> gives an insightful analysis of working in a VC funded company).

Exit strategies (extracting the profit from investments made in a company) used by VCs include selling startups to a listed company (reasons for large companies to buy startups range

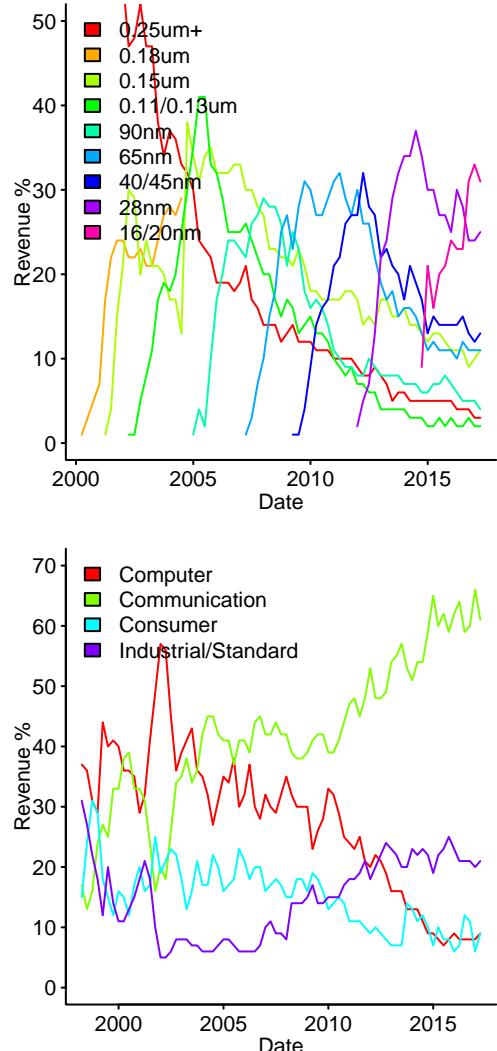


Figure 4.25: TSMC revenue from wafer production, as a percentage of total revenue, at various line widths. Data from TSMC.<sup>1367</sup> code

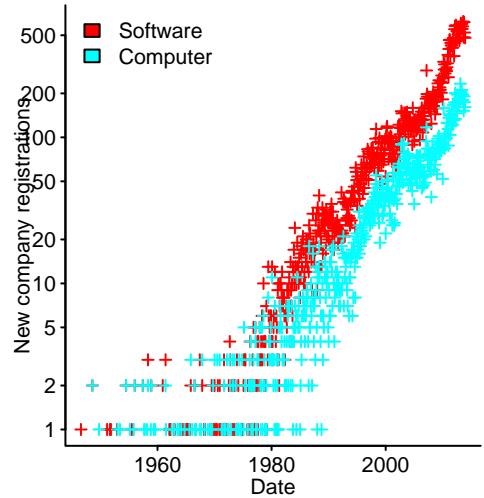
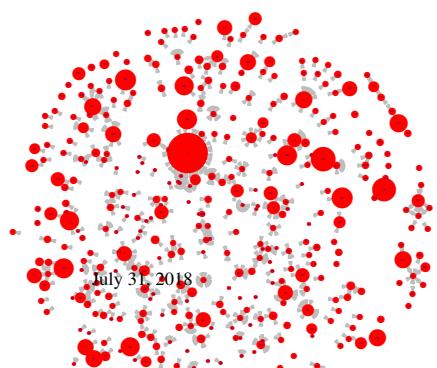


Figure 4.26: Number of new UK companies registered each month, whose SIC description includes the word software or computer (case not significant). Data extracted from OpenCorporates.<sup>1035</sup> code



<sup>viii</sup> Groups of students decided which companies to interview and so some clustering is the result of students using convenience sampling \text{---} email conversation with authors.

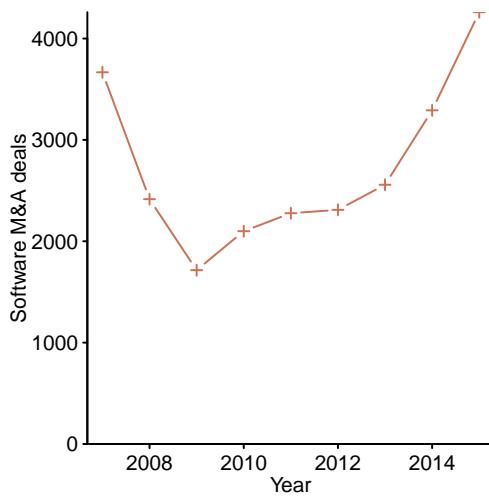


Figure 4.28: Reported worldwide software industry Mergers and Acquisitions (M&A). Data from Solganick.<sup>1278</sup> code

from acquiring people who have specific skills,<sup>1224</sup> removing potential future competitors and using the acquired company to generate income; see Figure 4.28) and an IPO (i.e., having the company's shares publically traded on a stock exchange; between 2011-15 the number of software company IPOs was in the teens and for IT services and consulting companies the number was in single digits<sup>1130</sup>). Venture capitalists are typically paid a 2% management fee on committed capital and a 20% profit-sharing structure; the VCs are making money, while those who invested in VCs over the last 20-years could have received greater returns by investing in a basket of stocks in the public exchanges.<sup>975</sup>

Red Queen data promised soon...???

#### 4.4.2 Cooperative competition

In ecosystems driven by strong network effects, there can be substantial benefits in cooperating with competitors; the cost of losing the war to set a new product standard can be very high.<sup>1232</sup>

Various standards' bodies, organizations, consortia and groups have been formed to document agreed-upon specifications. Committee members are often required to notify other members of any patents that have that impinge on the specification being agreed to; organizations that fail to reveal patents they hold and subsequently attempt to extract royalties from companies implementing the agreed specification may receive large fines and be required to license their patents under reasonable terms.<sup>416</sup>

Software systems need to be able to communicate with each other, agreed communication protocols are required. The communication protocols may be decided by the dominant player (e.g., Microsoft with its Server protocol specifications<sup>1467</sup>), by the evolution of basic communication between a few systems to something more complicated involving many systems, or by interested parties meeting together to produce an agreed specification.

A study by Simcoe<sup>1253</sup> investigated the production of communication specifications (i.e., RFCs) by the Internet Engineering Task Force (IETF) between 1993 and 2003 (the formative years of the Internet). Figure 4.29 shows that the time taken to produce an RFC having the status of a Standard increased as the percentage of commercial membership of the respective committee increased, but there was no such increase for RFCs not having the status of a standard; Simcoe proposed that the delay was caused by conflicts between vendors jockeying for commercial advantage...

### 4.5 Career ecosystems

Companies operating in knowledge-based ecosystems are becoming social factories,<sup>224</sup> involving themselves in employee's life as a means of increasing the cognitive effort they capture; free meals and laundry service are not perks, they are a means of bringing employees together to learn and share ideas by reducing opportunities to mix in non-employee social circles (and creating life norms that restrict the number of alternative employers, i.e., a means of reducing knowledge loss and spillover). The purpose of industrializing bohemia is to create an environment where knowledge-based workers feel happy and fulfilled spending their cognitive capital striving to achieve the goals set by their managers.

Company ecosystems that continue to be based around work-life separation also use software systems, and offer software related jobs. In these ecosystems the focus may be on the cognitive efforts of people working on non-software activities, with those working in software related activities having to fit in.

With so many benefits on offer to workers in knowledge-based ecosystems, employers need a mechanism for detecting free-riders; employees have to signal hedonic involvement, e.g., by working long hours.<sup>807</sup>

Two ethnographic studies of engineering culture are Kunda<sup>793</sup> of a large high-tech company in the mid-1980s, and Ross<sup>1175</sup> of an internet startup that had just IPO'd.

Higher education used to serve as a signalling system,<sup>1285</sup> for employers looking to recruit people at the start of their professional careers (e.g., high cognitive firepower was required to gain a university degree). However, some governments have a policy of encouraging a

significant percentage of students to study for a higher education degree, diluting university qualifications to being an indication of not below average IQ. By taking an active interest in the employability of graduates with a degree in a computing related subject,<sup>1228</sup> governments are showing signs of suffering from cargo-cult syndrome. Knowledge-based businesses want employees who can walk-the-talk, not drones who can hum a few tunes.

Human capital theory... self-exploitation... geekploitation...

Software companies employ people to perform a variety of jobs, including management,<sup>822</sup> sales, marketing, engineering, Q/A, customer support, internal support staff (e.g., secretarial). Figure 3.1 shows that even in a software intensive organization only around 87% of revenue is spent on non-software development activities.

What activities do careers in software development involve and how many people can reasonably claim to working in software development?

Census information and government employment statistics are sources of data covering many people likely to have a reasonable claim to be a software developer. However, this data may include jobs that are associated with software development. A study by Gilchrist and Weber<sup>498</sup> investigated the number of employed computer personnel in the US in 1970. The data for what was then known as *automatic data processing* included keypunching and computer operations personnel; approximately 30% of the 810,330<sup>ix</sup> people appear to have a claim to be software developers (see [ecosystems/50790641-II.R](#)). This data does excludes software development classified under R&D.

Regional pay differences... Estimating sizes of communities. [ecosystems/MSA\\_M2016\\_CM.R](#)  
bls occupational Employment Statistics [oes/](#), [O\\_Net](#)

? also see Figure 5.41

?

Hedonism driven software development ecosystems exist in various forms, including:

- Open source projects that do not pay any developers for contributing time and effort (a few projects include both paid and unpaid developers),

?

?

?, ?, ?

- advancement within academic ecosystems is driven by reputation, and writing software may be a component of research project; recognition is not always bestowed by those using software written by other academics.<sup>644,980</sup> Projects sometimes create home-brewed super-computers, e.g., [SETI@home](#)...

- the Maker community<sup>7</sup> builds hardware, which often requires software to drive it...

The sector economic model groups human commercial activities into at least three sectors:<sup>742</sup> the primary sector produces or extracts raw materials, e.g., agriculture, fishing and mining, the secondary sector processes raw materials, e.g., manufacturing and construction, and the tertiary sector provides services. The production of intellectual capital is sometimes referred to as the quaternary sector. Figure 4.30 shows the percentage of the US workforce employed in the three sectors over the last 160 years (plus government employment).

Figure 4.31 shows...

?

?, ?

?

?

<sup>ix</sup> 127,491 working for the Federal government, 27,839 in state government extrapolated from data on 36 states and estimated 655,000 in private establishments.

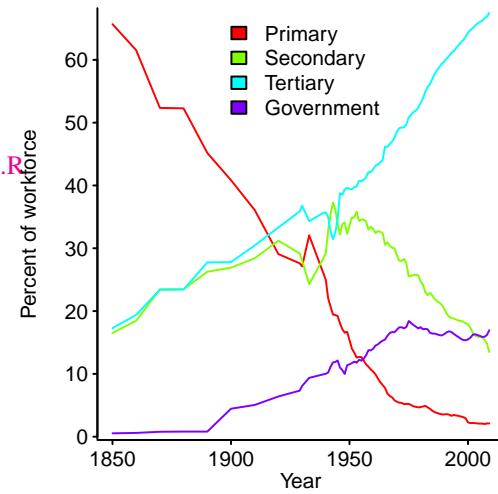


Figure 4.30: Percentage of employment by US industry sector 1850-2009. Data kindly provided by Kossik.<sup>690</sup>  
code

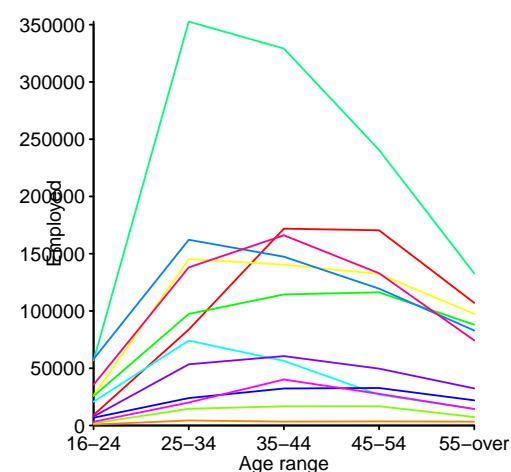


Figure 4.31: Number of people working in the 12 computer occupation codes assigned by the U.S. census bureau during 2014, stratified by ages bands (the "Software developers, applications and system software" code contains the largest percentage; see code for the identity of other occupation codes). Data from Beckhusen.<sup>116</sup> code

### 4.5.1 Career progression

The concept of career progression comes from an era where many industries were stable for long enough for career paths within them to become established.

Computers have created many new industries, but the stability needed to establish industry specific career paths has not existed. The general engineering career path is various levels of engineering seniority involving increasing management responsibilities, potentially followed by various levels of management.

Companies have an interest in employees progressing from managing code to managing people, it is a means of receiving greater value from a person based on their experience of the business. Employees may decide that they prefer to continue primarily working on technical issues, rather than people-issues, or may embrace a move to management. Simulations<sup>1095</sup> have found some truth to the Peter principle: “Every new member in a hierarchical organization climbs the hierarchy until they reach their level of maximum incompetence”.

A study by Jamtveit, Jettestuen and Mathiesen<sup>681</sup> investigated academic research units and found that the ratio of support staff to academic staff was fitted by a power law having an exponent of around 1.30 (a hierarchical management model of one administrator per three sub-units produces an exponent of 1.26).

In our changing world, does the concept of career progression make any sense? Some software companies have established career progression ladders for their employees;<sup>1289</sup> Gruber<sup>560</sup> gives a detailed discussion of the issues for IT careers.

If developers cannot be retained on a project, new ones need to be recruited. Computing related work has a reputation there being more vacancies than there are qualified people to fill them.<sup>7</sup> This means that jobs are sometimes filled by staff who do not appear as qualified as managers would like, either in application-domain knowledge or programming skill.

People change, the companies they work for change and the software ecosystem changes. There are plenty of opportunities for people to change jobs and the culture of information technology is one of high staff turnover compared to other industries<sup>7</sup> (with reported annual turnover rates of 25% to 35% in Fortune 500 companies).

In a fast changing knowledge-based industry there is a real risk of skills becoming obsolete. Options available to those whose skills are becoming obsolete include remaining in a job that looks like it will continue to require the skill for a long time, or learning new skills.

Employee turnover... people who prefer to write software, people who are happy to maintain software, people who want to manage... *Turnover Rates/...*

emailed for data... ???

Human capital theory suggests there is a strong connection between a person’s salary and time spent on the job at a company, i.e., the training and experience gained over time increases the likelihood that a person could get a higher paying job elsewhere; it is in a company’s interest to increase employee pay over time<sup>114</sup> (one study<sup>1266</sup> of 2,251 IT professionals in Singapore found a linear increase in salary over time).

A study by Joseph, Boh, Ang and Slaughter<sup>722</sup> analyzed BLS data to find the sequence of job held by people working in IT... extract data todo...

Implementing a new software is seen, by some developers, as being much more interesting and rewarding than maintaining existing software.<sup>7</sup> It is common for the members of the original software to move on to other projects once the one they are working on is initially completed<sup>7</sup> A study by Couger and Colter<sup>296</sup> investigated approaches to motivating developers working on maintenance activities; they identified the following factors:

- the motivating potential of the job: based on skill variety required, the degree to which the job requires completion as a whole (task identity), the impact of the job on others (task significance), degree of freedom in scheduling and performing the job, and feedback from the job (used to calculate a *Motivating Potential Score*, MPS),
- a person’s need for personal accomplishment, to be stimulated and challenged; the term *growth need strength* (GNS) was used to describe this.

The research provided support for the claim that MPS and GNS could be measured and that jobs could be tailored, to some degree, to people. Management's role was to organize the work that needed to be done to balance the MPS of jobs against the GNS of the staff available.

Computing jobs data<sup>7</sup> emailed...

The Occupational Employment Statistics published by the BLS in the US and the Labour Force Survey published by the ONS in the UK are country averages over everybody who files a tax return... [oes/ salary/](#)

In some US states, government employee salaries are considered to be public information, e.g., California (see [ecosystems/transparentcalifornia.R](#)); a few companies publish employee salaries, so called *Open salaries*; the information may be available to company employees only (e.g., ...), or it may be public (e.g., Buffer<sup>688</sup>).

If the male/female cognitive ability distribution seen in Figure 2.4 carries over to software competencies, then those seeking to attract more women into software engineering, and engineering in general, should be targeting the more populous middle competence band and not the high-fliers. The volume market for those seeking to promote female equality is incompetence; a company cannot be considered to be gender neutral until incompetent women are just as likely to be offered jobs as incompetent men.

There are gender differences in the pay of IT occupations<sup>1005</sup>...

Some developers are more strongly motivated by enjoyment of doing what they do, than the money received for doing it. This lifestyle choice relies on the willingness of companies to tolerate workers who are unwilling to do work they don't enjoy, or alternating between high paying work that is not enjoyed and working for pleasure. Freelancing in tasks such as trying to earn bug bounties is only profitable for a few people (see Figure 4.14).

What are the timeframes of personal involvement in an Open Source project?... TODO email... ??



Figure 4.32: Payer and payee countries of bug bounties (total value over ???). Data from hackerone.<sup>570</sup> [code](#)

## 4.6 Product ecosystems

A successful product in the market has a more complex support and dependency ecosystem than that of a bespoke system. The shift from one to many customers dilutes the influence of individual customers and vendors assume the risk of investing in any maintenance and enhancements; there is also greater opportunity for third-parties to become involved (e.g., sell add-on products and services).

Over time, customers' work-flow molds itself around the workings of a software system; an organization's established way of doing things evolves to take account of the behavior of the software it uses; staff training is a sunk cost. The cost of changing established practices, real or imaginary, creates a moat that makes it more difficult for customers to switch to competing products and is also a ball-and-chain that restricts the product updates to those that do not require customers to change; at some point the profit potential of new customers may outweigh that of existing customers, and an updated product requires existing customers to make an investment in adapting to the new release.

A regular release schedule allows the vendor and existing customers to plan ahead. The release schedule and version of some systems is sufficiently stable that a reasonably accurate regression model can be fitted<sup>1394</sup> (see [regression/release\\_info/cocoon\\_mod.R](#)). Issues, such as security vulnerabilities and pressure from the introduction of competitive products and in the case of open source projects availability of developers, can cause schedules to change...

Commercial products evolve when vendors believe that investing in updates is economically worthwhile. Updated versions of a product provide justification for asking customers to pay maintenance or upgrade fees, and in a competitive market work to maintain, or improve, market position; product updates also signal to potential customers that the vendor has not abandoned the product (unfavourable publicity about failings in an existing product that could deter potential new customers).

Public product version numbers are used to signal information such as which upgrades are available under a licensing agreement (e.g., updates with changes to the minor version number are included), and as a form a marketing to potential customers (who might view high numbers as a sign of maturity).

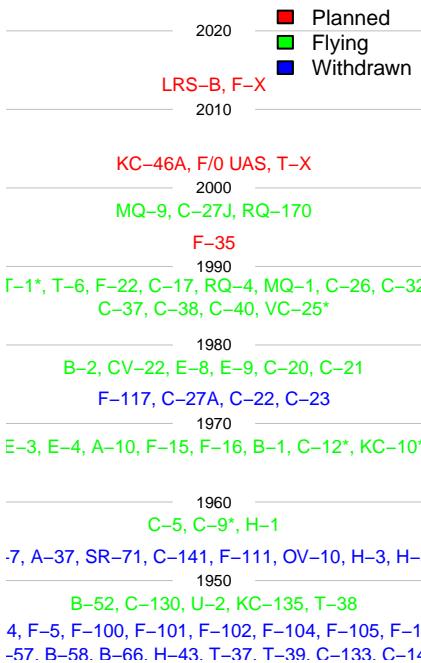


Figure 4.33: Decade in which newly designed US Air Force aircraft first flew, with colors indicating current operational status. Data from Echbeth et al.<sup>371</sup> code

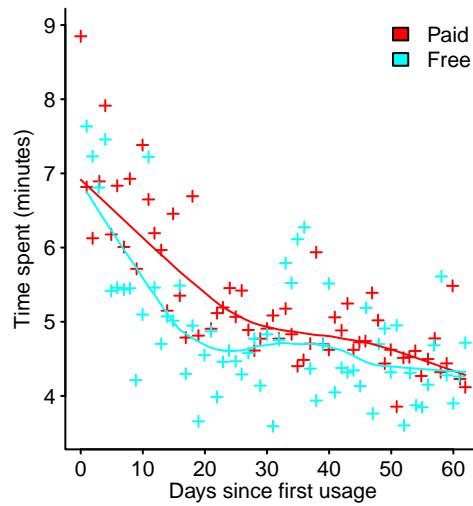


Figure 4.34: Daily minutes spent using an App, from Apple's AppStore, ... Data extracted from Ansar.<sup>45</sup> code

Time between application updates... emailed, pending...?

Software systems funded by hedonism continue to change for as long as those involved continue to enjoy the experience...

Customers want to get on with running their business and are happy to continue using existing software systems that meet their needs. The rate of product improvements will slow down over time, with product lifetimes of many decades becoming the norm. Figure 4.33 illustrates how the working life of jet-engined aircraft (invented in the same decade as computers) is increasing.

The days when customers bought their first computer to run an application are long gone. Except for specialist applications and general updates, there is rarely any reason for customers to invest in new computing hardware specifically to run the application. The characteristics of existing customer hardware is crucial because software that cannot be executed with a reasonable performance in the customer environment will not succeed. Figure 8.26 shows the variation in basic hardware capacity of desktop systems.

emailed...?

emailed for more data...?

In a few ecosystems the first release is everything and there is little ongoing market; Figure 4.34 shows the daily minutes spent using an App, installed from Apple's AppStore, against days since first used. This used to be the case when people paid for games software on their phones; a shift to in-game purchases created an ongoing relationship... emailed for data... .

Choice of App, rating of App, why do users pay for Apps, why do users stop using an App?...?

The regular significant improvement in Intel cpu performance (see Figure 4.8), starting in the last 1980s, became something that was factored into system development, e.g., future performance improvements were used as a reason for not investing too much effort tuning the performance of the next release.

#### 4.6.1 Product customization

Products are often customised for specific market segments. Reasons for this include simplification of product support, hardware resource constraints, market segmentation as a sales tool... .

Customization might occur during program start-up, when configuration information is read and used to control access to optional functionality, or at system build time, e.g., optional functionality is selected compile time, creating a customised program.

Each customized version of a product experiences its own evolutionary pressures and is a new potential source code program faults. patterns of coevolution<sup>1060</sup>...

A study by Rothberg, Dintzner, Ziegler and Lohmann<sup>1179</sup> investigated the number of optional Linux features supported by a given number of processor architectures. Table 4.1 shows that around 33% of features were supported by one architecture and 65% by almost all architectures.

The official Linux kernel distribution does not include all variants that exist in shipped products,<sup>599</sup> while manufacturers may make the source code of their changes publically available they either do not submit the changes to become part of the mainline distribution or their submissions are not accepted into the official distribution (it would be a heavy burden for the official Linux kernel distribution to include every modification made by a vendor shipping a modified kernel).

#### 4.6.2 Maintenance

In a stable ecosystem, by definition, most of the available resources are invested in maintaining of the current state of affairs; if significant resources were invested in change, the ecosystem could not be said to be stable... .

Version	1	2	3	All-3	All-2	All-1	All
<b>2.6.39</b>	3,989	182	50	2,293	944	1,189	2,617
<b>3.0</b>	3,990	183	53	2,345	968	1,211	2,637
<b>3.1</b>	4,026	184	52	2,440	968	1,155	2,667
<b>3.2</b>	4,028	181	57	1	2,788	512	4,054
<b>3.3</b>	4,077	180	51	1	2,837	512	4,133
<b>3.4</b>	4,087	183	51	1	2,907	520	4,184
<b>3.5</b>	4,129	179	50	2	3,001	520	4,265
<b>3.6</b>	4,158	184	51	2	3,098	527	4,298
<b>3.7</b>	4,139	183	50	1	3,173	539	4,384
<b>3.8</b>	4,148	178	35	3	3,269	548	4,399
<b>3.9</b>	4,269	177	36	3	3,403	581	4,413
<b>3.10</b>	4,280	173	35	3	3,447	577	4,460
<b>3.11</b>	4,270	178	33	2	0	0	8,654

Table 4.1: Number of Linux features shared between (i.e., supported) a given number of architectures (top row); All denotes all supported architectures, which starts at 24 and increases to 30. Data from Rothberg et al.<sup>1179</sup>

Once up and running, some software systems become crucial assets for operating a business, and so companies have no choice but to pay whatever it takes to keep them running. From the vendors' perspective, maintenance is the least glamorous, but often the most profitable aspect of software systems; companies sometimes underbid on to win a contract and make their profit on maintenance activities (see Chapter 5...).

A company that has been responsible for maintaining software for a customer is in the best position to estimate actual costs and the price the customer is willing to continue paying for maintenance (see Figure 3.4). Without detailed maintenance cost information other companies are unlikely to be willing to bid to take over maintenance of an existing system<sup>140</sup> (at least unless the customer is willing to underwrite their risk).

What are the main activities that occur during software maintenance? Perhaps the most widely cited breakdown of activities are those listed in a paper<sup>848</sup> published 38 years ago (obtained by analysing the responses given by 69 maintenance managers answering a questionnaire containing 50 questions, not the most reliable way of obtaining information)... Nosek & Palvia questionnaire results from 1990... More recent studies<sup>1203</sup> have found very different results (which is to be expected given the small sample size used in the earlier work). Software is maintained in response to customer demand and so the activities will be driven by this demand, e.g., very large systems in a relatively stable market<sup>963</sup> will have a different change profile than smaller systems in sold into a rapidly changing market.

The issues around fixing reported faults during maintenance are discussed in Chapter 6.

A study by Dunn<sup>367</sup> investigated the development and maintenance costs (total over the first five years) of 158 software systems from IBM. The systems varied in size from 34 to 44,070 man-hours of development effort and involved from 21 to 78,121 man-hours of maintenance. Figure 4.35 shows the ratio of development to five-year maintenance costs. The data is for systems at a single point in time, 5-years. Modeling, using expected system lifetime, finds that the mean total maintenance to development cost ratio is less than one (see [ecosystems/maint-dev-ratio.R](#)). The correlation between development and maintenance man-hours is 0.5 (0.38-0.63 is the 95% confidence interval, see [economics/maint-dev-cost-cor.R](#)).

emailed<sup>2</sup>

extracted...?

?

A study by Tamai and Torimitsu<sup>1336</sup> obtained data on the lifespan of 95 software systems (appears to be mostly in-house systems). Figure 4.36 shows the number of systems surviving for at least a given number of years and a fit of two equations (with  $a$ ,  $b$  and  $c$  as constants):  $systems = a \cdot e^{b \cdot years}$  (green) and  $systems = a \cdot e^{b \cdot years + c \cdot years^2}$  (blue).

System half-life for this Japanese corporate mainframe data from 1991 is around 5-years. To what extent is this data applicable today? A shorter lifecycle may simply mean that more maintenance is done in less time, i.e., total maintenance costs are the same, or if roughly the same is done maintenance costs will be lower...

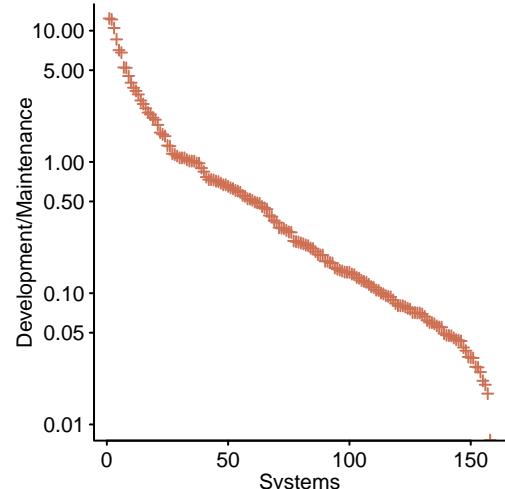


Figure 4.35: Ratio of development costs to total five-year maintenance costs for 158 IBM software systems sorted by size; curve is a beta distribution fitted to the data (in red). Data from Dunn.<sup>367</sup> [code](#)

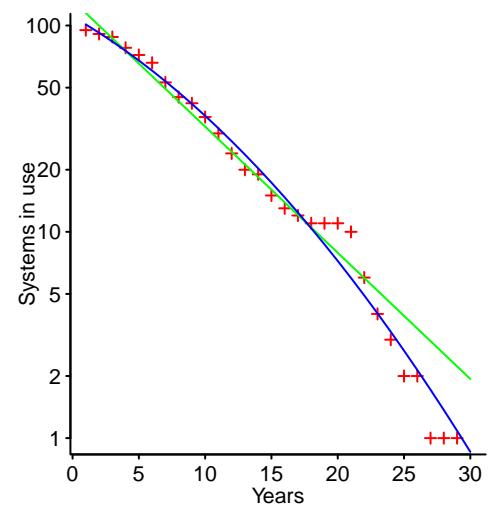


Figure 4.36: Number of software systems surviving to a given number of years and exponential equation fits. Data from Tamai.<sup>1336</sup> [code](#)

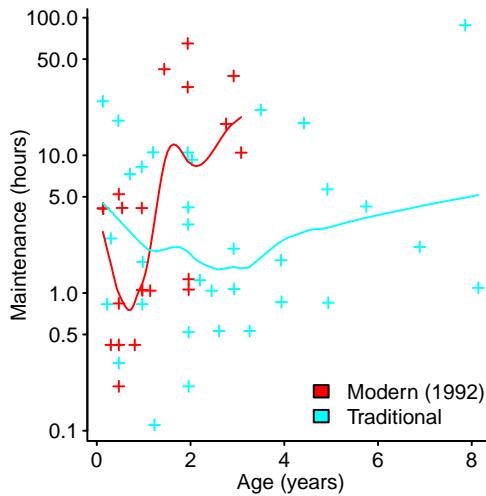


Figure 4.37: Age of systems, developed using one of two methodologies, and corresponding monthly maintenance time, along with loess fits. Data extracted from Dekleva.<sup>338</sup> code

+ A study by Dekleva<sup>338</sup> investigated the average monthly maintenance effort (in hours) spent on products developed using traditional and modern methods (from a 1992 perspective). Figure 4.37 shows the age of systems and the corresponding time spent on monthly maintenance. The loess fits suggest that average monthly maintenance for projects developed using a *traditional* method does not vary much with age; it is not known whether the maintenance level for *modern* projects maintains its higher level or declines in later years.

Product lifetime data...

+ A study by Baysal, Kononenko, Holmes and Godfrey<sup>111</sup> tracked the progress of 34,535 patches submitted through the WebKit and Mozilla Firefox code review process between April 2011 and December 2012. Figure 4.38 shows the percentage of patches (as a percentage of submitted patches) being moved between various code review states in WebKit. TODO Blink data...

The extent to which developer fault handling review behavior depends on the company who submitted the patch... acceptance/rejection time for apple/google/other?

In safety critical applications the impact of changes during maintenance has to be thought through,<sup>321</sup> this issue is discussed in Chapter 6.

### 4.6.3 Forking

Forking a system involves creating a duplicate of the directories and files needed to build the software and then working on the duplicated system as if it were separate, or independent, of the original. Forks occur for a variety of reasons (e.g., a company buying a license), but all involve the creation of a new development path that is separate from the original.

All it takes to fork a system is access+rights to the necessary source files and one or more developers with the inclination and ability to continue developing a project along a different path. In the open source world forking is sometimes seen as a social issue, e.g., a leadership issue over the direction of a project.

The *Fork* button on the main page of every project on the GitHub website is intended as a mechanism for developers, who need not be known to those involved in a project, to easily copy a project from which to learn and perhaps make changes; possibly submitting any changes back to the original project, a process known as *fork and pull*. As of October 2013, there were 2,090,423 forks of the 2,253,893 non-forked repositories on GitHub.<sup>687</sup>

A study by Robles and González-Barahona,<sup>1167</sup> in 2011, attempted to identify all known significant forks; they identified 220 forked projects, based on a search of Wikipedia articles, followed by manual checking. Figure 4.39 suggests that after an initial spurt, the number of forks has not been growing at the same rate as the growth of open source projects.

The BSD family of operating systems arose from forking during the early years of their development and have evolved as separate but closely related projects since the early-mid 1990s. Figure 9.21 suggests how a few developers working on multiple projects communicate bug fixes between them.

A study by Ray<sup>1144</sup> investigated the extent to which code created in one of NetBSD, OpenBSD or FreeBSD was ported to either of the other two versions, over a period of 18 years. Ported code not only originated in the most recently written code, but was taken from versions released many years earlier. Figure 4.40 shows the contribution made by 14 versions of NetBSD (versions are denoted by stepping through the colors of the rainbow) to 31 versions of OpenBSD; the contribution is measured as percentage of lines contained in all the lines changed in a given version.

Forked projects provide interesting data for research because they provide a mechanism for comparing the characteristics closely similar systems.

### 4.6.4 Product obsolescence

In a rapidly changing ecosystem (driven by a constant stream of dramatic changes in computing hardware), the incentive to maintain and support existing products can quickly disappear. It may be more profitable to create new products that are incompatible with current products, current products may not have sold well enough to make further investment worthwhile,

or the market itself may shrink to the point where it is not economically viable to continue operating in it.

Hardware products wear out, or break, and vendors can profit from replacement sales. Software does not wear out, there are no replacement sales. While software does not wear out, it does depend on an ecosystem and changes to third-party components that are depended upon can cause programs to malfunction, or even fail to execute. Changes in the outside world can cause unchanged software to break.

Some vendors continue to support and maintain products for a period after the next generation of products has been launched, providing an opportunity for customers to decide when to migrate to a later version. [data?](#)...

Some people create and support their own Linux distribution, often based off of one of the major distributions (e.g., Ubuntu was originally derived from Debian). Lundqvist and Rodic<sup>41</sup> recorded the life and death of these distributions and Figure 4.41 shows the survival curve, based on the parent distribution.

A study by Caneill and Zacchiroli<sup>216</sup> investigated packages in the Debian distribution. Figure 4.42 shows the survival curve of packages included in the standard Debian distributions.

The desktop market growth to dominance of Wintel<sup>x</sup> reduced platform diversity (i.e., the alternatives went out of business), which reduced the incentives for companies to continue to support a diverse range of platforms. The incentives changed to specializing in servicing the large market share held by the dominant platform...

File formats can have very long lifetimes; once a file is created using a given version of a format it may exist unchanged and continue to be used for many years. A typical vendor strategy is to continue supporting the ability to read older formats, but only supporting the writing of more recent formats. A study by Jackson<sup>677</sup> investigated the files created using a given file format available on websites having a UK web domain between 1996 and 2010. Figure 4.43 shows the total number of pdf files created using a given version of the pdf specification.

The market share of the latest version of a software system typically grows to a peak, before declining as newer versions are released. Villard<sup>1410</sup> tracked the Android version usage over time. Figure 4.44 shows the percentage share of the Android market held by various releases, based on days since launch of each release.

?

## 4.6.5 Documentation

The term *documentation* might be applied to anything, requirements, specifications, code documentation, comments in code, testing procedures, bug reports and user manuals; source code is sometimes referred to as its own documentation or as self-documenting.

The following are possible incentives for writing non-code documentation:

- fulfil a contract requirement. This requirement may have appeared in other contracts used by the customer and nobody is willing to remove it; perhaps customer management believes that while they do not understand code, they will understand prose,
- a costly display signalling commitment, e.g., management wants everybody to know that the project will be around for a long time or is important,
- an investment intended to provide a worthwhile return by reducing the time/cost of learning for future project hires... and the data for this...
- as marketing material for an individual offering consultancy or training services. A high quality book or manual may reduce the size of the potential pool of clients, but those remaining are made aware of a potential high quality supplier.

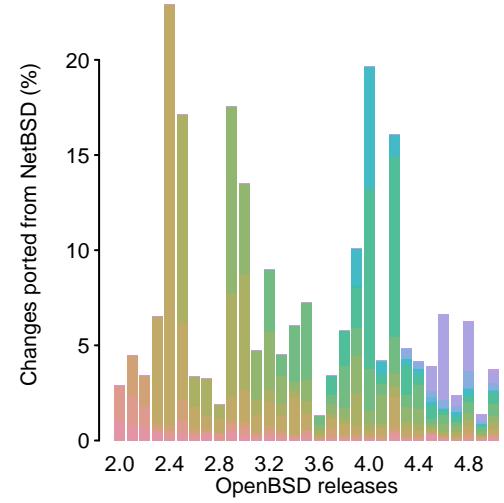


Figure 4.40: Percentage of code ported from NetBSD to various versions of OpenBSD, broken down by version of NetBSD in which it first occurred (denoted by incrementally changing color). Data kindly provided by Ray.<sup>1144</sup>

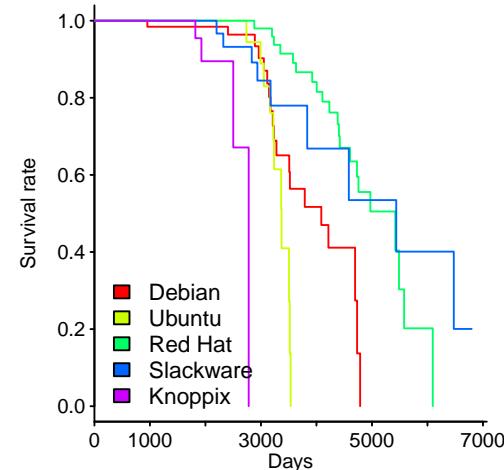


Figure 4.41: Survival curve of Linux distributions derived from five widely-used parent distributions (identified in legend). Data from Lundqvist et al.<sup>41</sup> [code](#)

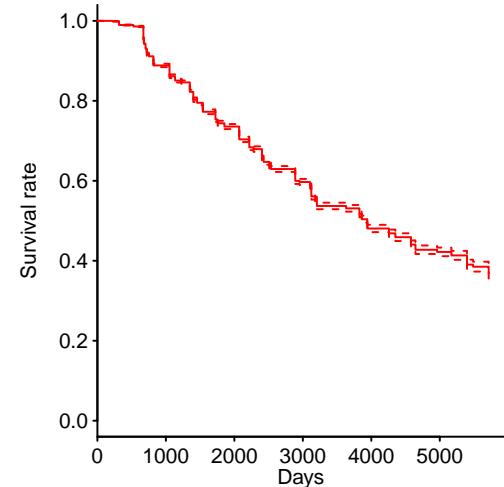
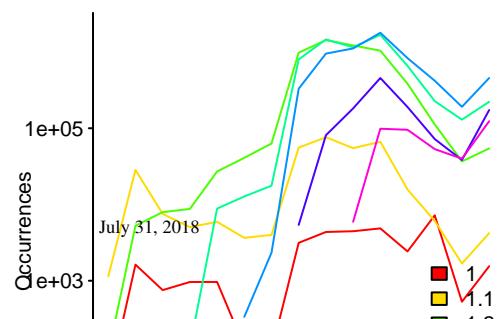


Figure 4.42: Survival curve for packages included in the standard Debian distribution. Data from Caneill et al.<sup>216</sup> [code](#)



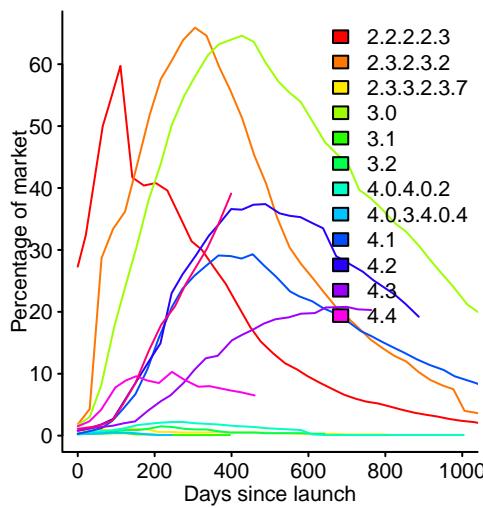


Figure 4.44: Percentage share of total Android market at days since launch for various versions of Android. Data from Villard.<sup>1410</sup> [code](#)

Development projects which derive income from consultancy and training have an incentive to minimise the amount of documentation. Knowledge of how a software system works has a value and producing documentation reduces that value.

The existence of documentation can be a liability, in that courts have considered vendors liable for breach of warranty when differences exist between product behavior and the behavior specified in documentation.<sup>732</sup>

How much are documents changed over time, what gets changed over time... Figure 4.45 shows...

Vendors that support an API have to document the calls available and provide examples, if they expect developers to make use of the services provided. APIs evolve and the documentation has to be kept up to date.<sup>1245</sup> The growth of popular APIs, over time, can result in an enormous collection of documentation, e.g., the 130+ documents for the Microsoft Server protocol specifications<sup>1467</sup> contain over 16 thousand pages (see Figure 8.23).

Which parts of an API document are read most often?... fig 5 emailed...

Issues around interpreting the contents of documentation are covered in Chapter 6...

Using topic models to track changes over time...?

?

## 4.7 Software development ecosystems

The early computers were delivered without software, customers were expected to write their own. By supplying basic software functionality manufacturers decreased the cost of using a computer, which increased the number of potential customers; Figure 4.11 shows how the amount of code included with the computers from one manufacturer increased over time.

High level programming languages (e.g., Fortran in 1954<sup>75</sup> and Cobol in 1961<sup>123</sup>) were created with the intent of reducing the cost of building software for different computers. The cost reduction came through portability of the software to a wider variety of computers and portability of the programming skills that had been gained on one kind of computer to different computers (e.g., removing the need to learn the assembly language for each new machine).

The workflow of software development influences the characteristics considered important in tools. For instance, during the first few decades, interaction with computers was often via batch processing rather than one-line access,<sup>509</sup> jobs (i.e., a sequence of commands to execute) were submitted and queued to be run. Developers might only get to complete one or two edit/compile/execute cycles per day. In this environment high quality compiler syntax error recovery is essential; having an executable for an error corrected version of the submitted source was more useful than just having a list of errors found, the generated binary could often be patched.<sup>xi</sup> Borland's Turbo-Pascal stopped compiling at the first error it encountered, dropping the user into an editor with the cursor located at the point the error was detected. Compilation was so fast, in an interactive environment on a personal computer, developers loved using it.

Hardware vendors were the only suppliers of software development tools until suppliers of third-party tools had access to the necessary hardware and the potential customer base became large enough to make it commercially attractive.<sup>xii</sup>

When developer tools were sold for profit, vendor had an incentive to keep both customer and tool user happy.<sup>xiii</sup> Open source has significantly reduced the number of customers for some tools, while maintaining similar numbers of users for these tools. For instance, the few companies developing their own unique CPU might pay for a code generator, for an open source compiler, to be written for this processor, with users of the compiler not paying anything. The focus of commercial companies is on their customers, not the people who use their software.

What does a software developer ecosystem contain, how do they interact and what are the patterns of evolution? and the data...

<sup>xi</sup> Your author once worked on a compiler project funded with the aim of generating code 60% smaller than the current compiler. Developers hated this new compiler because it generated very little redundant code; the redundant code generated by the previous compiler was useful because it could be used to hold patches.

<sup>xii</sup> Early computers installed in some universities resulted in the first open source software.

<sup>xiii</sup> Customers pay money, users use; a developer can be both the customer and the user.

Established ways of doing things, programming languages, libraries, tools (e.g., build systems and source code control, available third-party libraries), question/answer sites, forges and a collection vendors making a living from supporting developer activities (e.g., offering training, third-party libraries and compilers)...

The creation of major new ecosystems provides an opportunity for new languages and tools to gain a significant following... New languages only spread to achieve a significant market share when... cryptocurrency smart contracts?

Changes to compilers to support more aggressive optimizations, changes to the supported language... vendor extensions, new standard features... Figure 4.46 shows the growth in gcc compiler flags and options...

Survival curve of gcc support for various processors...

What do developers spend their time doing? ... 20% of time spent on build issues, 11% of time spent maintaining build... Startup cost or one time change cost for the build...?

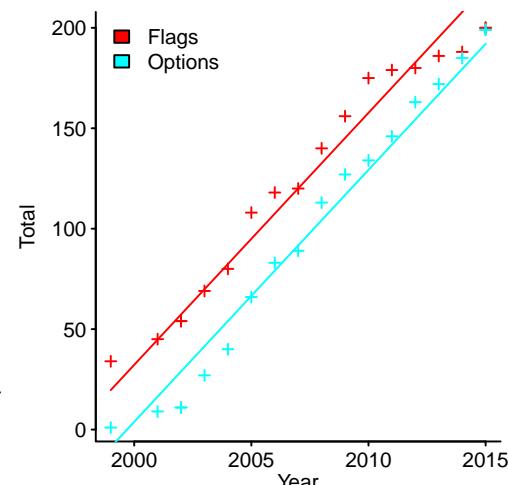


Figure 4.46: Number of gcc compiler flags and options over time, with fitted regression models. Data from Fursin et al.<sup>467</sup> [code](#)

## 4.7.1 Programming languages

Humans appear to have an innate desire to create languages,... writers of science fiction create elaborate and linguistically viable languages for aliens to speak<sup>1174</sup>...

The first published specification of a high-level programming language, Plankalkül, was in 1949;<sup>105,1503</sup> the first published specification for Fortran (which remains widely used) was in 1954.<sup>658</sup> A book published in 1959<sup>532</sup> lists around 25 compilers<sup>xiv</sup> [ecosystems/Grabbe\\_59.txt](#) and it was already being noted in 1963<sup>587</sup> that creating a programming language was a fashionable activity. A 1976 report<sup>437</sup> estimated that at least 450 general-purpose languages and dialects were currently in use in the US DoD.

Figure 4.47 shows the number of new programming languages, per year, described in a published paper...

High-level languages did not immediately displace machine code for most software systems. A 1977 survey<sup>1294</sup> of programmers in the US Federal Government, found 45% with extensive work experience and 35% with moderate work experience of machine code. Developing software using early compilers could be time-consuming and labor-intensive; memory limits required compiling to be split into a sequence of passes (often half-a-dozen or more<sup>182</sup>) over various representations of the source, with the intermediate representations being output on paper-tape which was read back in as the input to the next pass (after reading the next compiler pass from the next paper-tape in the sequence), eventually generating assembler. Some compilers required that the computer have an attached drum-unit<sup>77</sup> (early form of hard-disk), which was used to store the intermediate forms (and increase sale of peripherals).

The forces driving programming language ecosystems include (recognizable ecosystems grow up around widely used languages, while other languages have to fit in where they can):

- the quantity of source contained in commercially supported applications written in that language. Companies need developers with some degree of familiarity with the language to maintain and enhance these applications,
- the number of developers who can use the language well enough to be employed, writing code in it...
- perception (i.e., not reality) within developer ecosystems of which languages are worth knowing, becoming popular or declining in usage/popularity. In a rapidly changing environment developers want a CV that lists experience in employable languages to make them attractive to employers,
- existing libraries and support tools available to users of the language...
- compilers available on a given platform. A language cannot be used if no compiler is available for it; the early use of C for embedded systems has resulted a C compilers being one of the first developer tools created for a new processor...

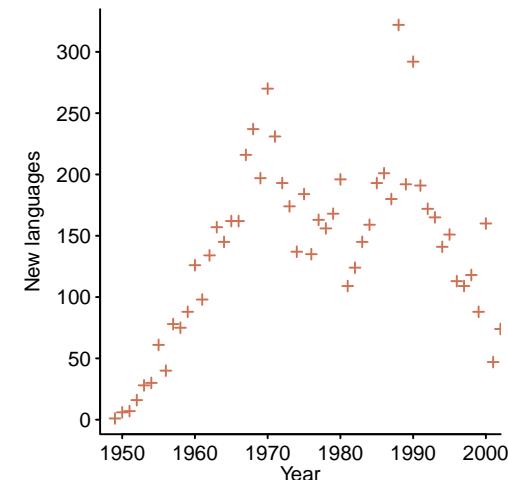


Figure 4.47: Number of new programming languages, per year, described in a published paper. Data from Pigott et al.<sup>1091</sup> [code](#)

<sup>xiv</sup> The term compiler was not always applied in the sense that is used today.

Other, one-off, reasons include local developer response to a country's strong trade barriers<sup>659</sup> ...

Many thousands of programming languages have been created, but only a handful have become widely used.<sup>xv</sup> The reason for the growth in popularity of a particular programming language within application ecosystems appears to be happenstance, e.g., Forth for embedded systems in Astronomy... Pascal remains popular in Russia... In Japan Cobol remains popular... different cultural forces in play...

The term *language popularity* suggests that the users of the language had some say in the decision to use the language and liked the language in some way. In practice developers often have no say in the choice of language and may have no positive (or negative) feelings towards the language. However, this term is in common use and there is nothing to be gained by using something different...

Sources of information on language usage include:

- Job adverts: Appearing to be willing to pay money for someone to use a language is not necessarily a good indicator of language usage; languages listed in job adverts have been chosen for a variety of reasons including: wanting to appear trendy in order to attract young developers, generating a smokescreen to confuse competitors and that knowledge of the language is required for the job advertised. The number of times a language is listed in a job advert is a measure of that language's perceived and actual popularity,

The UK news magazine Computer Weekly<sup>xvi</sup> publishes rankings of keywords appearing in the job adverts it contains.

Social media includes job postings and adverts for jobs. Figure 4.48 shows the number of monthly developer job related tweets that included a given language name... Recruitment agencies... emailed for data...

People read job adverts to help decide which language to learn to improve their chances of being offered a job.

- Quantity of existing code: this is a measure of total popularity since developers first started using the language. Most existing code is likely to be proprietary and not easily available to be measured. Open source is available to be measured... most only goes back under 20 years?... how old is most proprietary code?...

The rate at which new code is created... Figure 11.10

Figure 4.49 shows the number of languages used in a sample of 100,000 GitHub projects (make was not counted as a language).

- Number of books sold:<sup>943</sup> spending money on a book is an expression of intent to learn the language, which, if carried through, results in code being written in it (which will increase the quantity of existing code).

Publishers have a strong interest in new languages for which there are few, or no, existing books to learn from... so more interested in new languages that probably have a low usage ranking... [books-sold/](#)

- Miscellaneous available information, such as the number of times the name of a language is mentioned in books or question/answer sites. Suggestions that social media data be measured are inevitably made, but given that names of programming languages are often also names of other entities any such data would need to be carefully cleaned before it can be considered at all reliable.

Number of commercially available compilers as an indicator of market size (prior to the growth of open source)...

US Federal government surveys of its own usage: a 1981 survey<sup>281</sup> found most programs were written in Cobol, Fortran, Assembler, Algol and PL/1, a 1995 survey<sup>636</sup> of 148 million LOC in DOD weapon systems Ada represented 33%, the largest percentage of any language (C usage was 22%)...

---

<sup>xv</sup> It is said that there are two kinds of programming language: those that everybody complains about and those that nobody uses.

<sup>xvi</sup> Until the Internet era Computer Weekly and Computing were the primary UK public sources of job information. Publication of a weekly paper based version ceased in April 2011 and content now appears via the website [computerweekly.com](http://computerweekly.com)

Language popularity for mobile phone development...?

Programming languages used in embedded systems on a current project, vdcresearch survey... lang-pop-13-vdcresearch.jpg [embedded\\_survey/](#)

??

???

Published books are a source of serious discussion about programming languages going back many years and the Google has released the n-grams of the words appearing in the books it has scanned; <sup>942</sup> the unigram and bigrams for English were used in the following analysis.

Language names such as Fortran and Cobol are unlikely to be used in non-programming contexts, while names such as Java and Python have names that could be mentioned more often than the programming language. Single letter names, such as C, or names followed by non-alphabetic characters... The phrase *in C* also appears in books on music (i.e., the key signature of a piece of music) and the OCR process sometimes inserts spaces that probably did not exist in the original.

Java has multiple non-computer related uses and subtracting the estimated background usage suggests a language usage similar to that of *SQL*.

Multiple languages used within a project...?

A study by Savić, Ivanović, Budimac and Radovanović<sup>1201</sup> investigated the impact of a change of teaching language on student performance in practical sessions (Modula-2 to Java). Student performance, measured using marks assigned, was unchanging across the four practical sessions, as was mean score for each year (see [ecosystems/2016-sclit-uup.R](#) for details).

Some programming languages evolve, i.e., new features are added and existing features are removed. One consequence of language evolution is that there may be implementations of the language that do not support constructs contained in the source code used to build a program<sup>127</sup>...

?

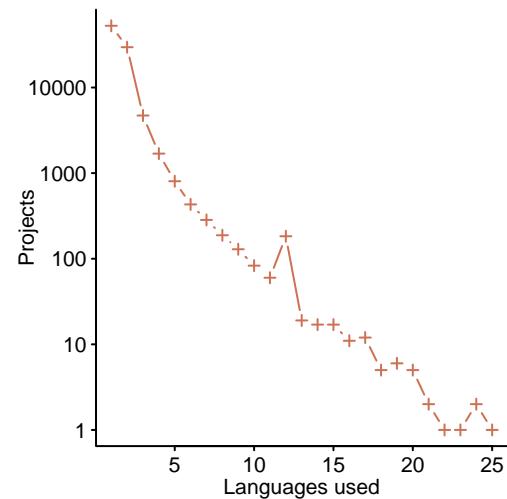


Figure 4.49: Number of projects making use of a given number of different languages in a sample of 100,000 GitHub projects. Data kindly supplied by Bissyande.<sup>145</sup> [code](#)

## 4.7.2 Libraries and packages

Libraries of commonly used subroutines were created for the first computers<sup>1454</sup> and computer vendors decided which functions were included in the libraries provided for customers.<sup>146</sup> Programming languages were often designed with specific uses in mind and, to ensure application portability, designers specified a minimum set of library functions that had to be supported, e.g., Fortran targeted engineering/scientific users and required support for trigonometric functions, Cobol targeted business users and required support for sorting.<sup>xvii</sup>

The decreasing cost of hardware and the availability of an operating system, in the form of Unix source code, allowed many companies to enter the minicomputer/workstation market. Vendor's attempts to differentiate themselves led to the Unix wars<sup>1195,1196</sup> of the 1980s (in the mid-90s platforms running a Unix-derived OS typically shipped with over a thousand C/C++ header files<sup>699</sup>).

Supporting a standard set of library functions that are often required by applications would reduce porting costs. The POSIX ISO standard was intended to enhance source code portability across operating systems<sup>1502</sup> by supporting a basic set of library functions and basic utility programs. POSIX became widely supported (in part because large organizations, such as the US Government required vendors to supply products that included POSIX support, although some implementations feel as-if they are only intended to tick a box for a tender process, rather than be used...).

A study by Atlidakis, Andrus, Geambasu, Mitropoulos and Nieh<sup>60</sup> investigated POSIX usage (1,177 functions) across Android 4.3 (1.1 million apps measured, 790 functions tracked out of 821 implemented) and Ubuntu 12.04 (71,199 packages measured, 1,085 functions tracked out of 1,115 implemented). Figure 4.50 shows how application usage of POSIX functions varies across even closely related operating systems (these numbers do not include calls to *ioctl* whose action is to effectively perform an implementation defined call).

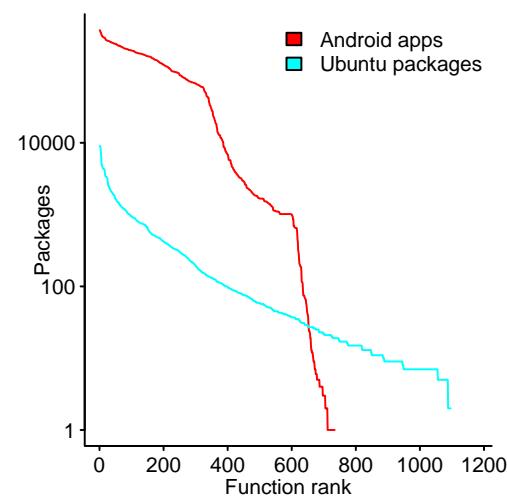


Figure 4.50: Ranked order of number of Android/Ubuntu (1.1 million apps)/(71,199 packages) linking to each supported POSIX function. Data from Atlidakis et al.<sup>60</sup> [code](#)

<sup>xvii</sup> The C Standard specifies support for some surprising functions, surprising until it is realised that they are needed to implement a C compiler, e.g., *strtoul*.

Linux came late to the Unix wars and emerged victorious. The Linux Standard Base<sup>xviii</sup> is intended to support a binary compatibility interface for application executables; this interface includes pseudo-file systems (e.g., `/proc`) that provide various kinds of system information. A study by Tsai, Jain, Abdul and Porter<sup>1365</sup> investigated use of the Linux API by Ubuntu 15.04 applications, including system calls and pseudo-file system usage... schema promised...

The internet made it possible to sustain an ecosystem of packages and libraries for popular programming languages, e.g., npm for Javascript and CRAN for R. In some cases a strong symbiotic relationship has formed which has influenced the uptake of new language revisions, e.g., one reason for the slowed uptake of Python 3 has been use of existing packages that have required the use of Python2...

The installation of a package may fail because of dependency conflicts between the packages already installed and this new package, e.g., installed package  $P_1$  may depend on package  $P_2$  having a version less than 2.0, while package  $P_3$  depends on package  $P_2$  having a version of at least 2.0. A study by Drobisz, Mens and Di Cosmo<sup>363</sup> investigated Debian package dependency conflicts. Figure 4.51 shows a survival curve of package lifetime and the time at which the first conflict occurs for a package.

Package dependency conflicts only become an issue if an attempt is made to install the packages with conflicting dependencies... A study by Decan, Mens and Claes<sup>329</sup> investigated how package dependencies in three ecosystems (npm, CRAN and RubyGems), changed over time.... [ecosystems/SANER-2017.R](#)

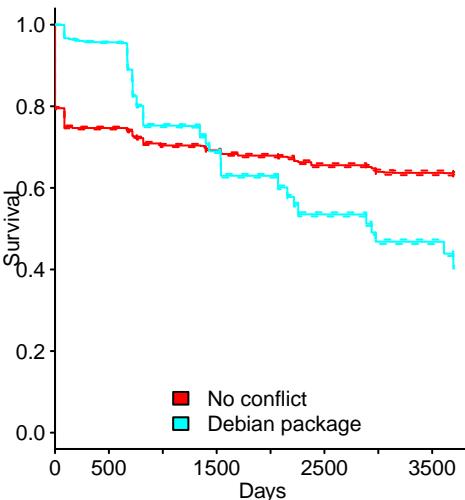


Figure 4.51: Survival curves for Debian package lifetime and for a package to contain its first dependency conflict. Data from Drobisz et al.<sup>363</sup> [code](#)

The requirements for a package to be included in an established repository may conflict with the workflow of a package under active development (where the package authors want the latest release to be rapidly available to potential users), for instance, some R packages under active development are available on GitHub (moving to CRAN when they become stable). There are dependencies between packages in the two repositories and dependency conflicts exist<sup>330</sup>...

Languages evolution can introduce incompatibilities in their API...<sup>791</sup>

?

?

The issue of compatibility of Eclipse third-party plug-ins (ETP) with the Eclipse SDK is discussed elsewhere (see Figure 11.71).

### 4.7.3 Licensing

Many different software licenses have been created, including a proliferation of different and sometimes incompatible open source licenses. While many people do not read end user license agreements (EULAs),<sup>898</sup> those responsible for public repositories and distributions sometimes restrict the software they make available to that licensed under a particular agreement (or licenses compatible with it), e.g., Debian distributions only contain software whose source is licensed under an Open Source license.

A study by German, Manabe and Inoue<sup>489</sup> investigated the use of licenses in the source code of programs in the Debian distribution (version 5.0.2). They found that 68.5% of files contained some form of license statement (median size of the license 1005 bytes). The median size of all files is 4633 bytes; the median size of files without a license 2137 bytes and the files with a license 5488.

How frequently do the licenses in source files change?<sup>7</sup> The Unity game development tool license for 2008 included 26 different open sources licenses, while Google's Chrome license in 2011 included 27 different open source licenses.

Licensing rights for data associated with software<sup>594</sup>...

???

<sup>xviii</sup> LSB 3.1 was published as an ISO Standard in 2006, however this Standard has not been updated to reflect later versions of the LSB. The Linux API has evolved.<sup>82</sup>

## 4.8 Evolution of source code

Software only changes when developers have an incentive to spend time making the changes. Incentives for developers to spend their time in this way include being paid and a desire to change the code to satisfy a personal need (e.g., refactor code in the belief that other developers will have a higher opinion of the code...).

If payment is involved, there is a customer, and the changes are intended to address customer needs (it can be very difficult to work out what the customer needs actually are and there may be as many opinions about these needs as there are people working to keep the customer happy).

Software systems growth, in lines of code, over time is a commonly seen characteristic; common reasons for growth include improvements to existing functionality and the addition of new functionality. Some systems grow at a consistent rate over many years, e.g., FreeBSD Figure 11.2 and the Linux kernel Figure 11.7, while others appear to have stopped adding lines, e.g., Figure 11.50 or grow sporadically, e.g., the Groovy compiler Figure 11.9.

Growth can also increase interdependencies between components; Figure 4.52 shows the relationship between the separate components of ANTLR over various releases.

The following are the primary factors influencing the evolution of source code characteristics:

- customer limited: not be enough customer demand (as measured by a willingness to pay) for it to be economically worthwhile updating existing functionality (to support changes in the world) or adding new functionality, e.g., new hardware requiring device drivers for a particular operating system,
- developer limited: bottlenecks in the development process that restrict the quantity of change per unit time. For instance, a limited number of people with the necessary skills, changes requiring sign-off by a handful of senior managers or increasing effort required to support a growing system leading to diminishing returns from adding more developers,
- some combination of customer and developer limits: variability of customer demand creates uncertainty about when and how much developer resource will be needed; resource scheduling... leading to a stop-go effect...

A consistent rate of growth suggests some degree of consistency in customer demand and the developer effort available to do the work...

Updating existing functionality can result in source code being deleted. The turnover of source code in Linux is evident in Figure 4.53, which shows the percentage of code in 130 releases that originated in earlier releases, and Figure 4.40 shows code shared between different releases of related BSD operating systems; Figure 11.66 shows the correlation between lines added/deleted for glibc, Figure 9.20 shows a Markov chain for file creation/modification/deletion in the Linux kernel...

While a surprisingly large percentage<sup>374</sup> of program features may never be used, they are implemented because future usage is assumed to be likely...

Removal of source from gcc that supports code generation for older processors over time...

shi2013.R or Volvo Masters...?

emailed for more data...?

Hardware capacity limits place upper bounds on the maximum size of programs that can be executed and incentives to continue working on an application limit the amount of code that gets added, modified or removed from it....

kbuild and linux commits per day/week... Time series of commits to the Linux source and configurations files...? Fig 3

Evolution of database frameworks...??

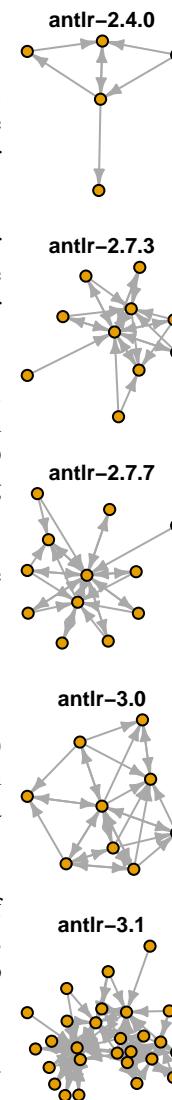


Figure 4.52: Dependencies between the Java packages in various versions of ANTLR. Data from Al-Mutawa.<sup>18</sup> code

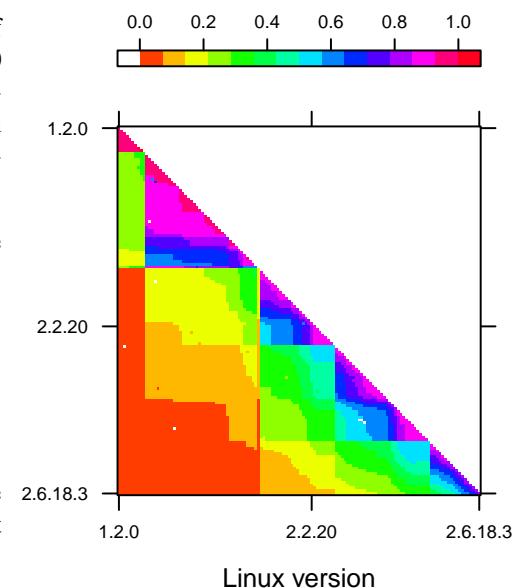


Figure 4.53: Fraction of source in 130 releases of Linux (x-axis) that originates in an earlier release (y-axis). Data extracted from png file kindly supplied by Matsushita.<sup>859</sup> code

### 4.8.1 Source code lifetime

The survival rate of source code is a crucial input value to any cost/benefit investment analysis. Termination of source code usage occurs at various levels of granularity, from the top-level of the application or library containing it ceasing to be used, to the deletion of individual lines of code.

A function or method is a self-contained unit of code and an obvious candidate for a cost/benefit analysis. Analyzing a sequence of lines, of code, is likely to be complicated because of the need to estimate the impact of the other locally associated lines (in many languages these would be contained within the same function).

A new function definition is about to be written and it is thought that at some future time it may need to be modified. If an investment,  $I$ , in extra work is made to receive a benefit,  $B$ , during each future modification, what is the relationship between  $I$  and  $B$ ?

Like all investments, the expected benefit has to be greater than the investment:

$$I < M_t B$$

where  $M_t$  is the total number of modifications of the function.

Let  $0 < s < 1$  be the probability that a function will be modified in the future and that once modified the likelihood of it being modified again remains unchanged; the expected number of modifications of a given function is then:

$$M_t = s + 2s^2 + 3s^3 + \dots + ns^n$$

where  $n$  is the maximum number of modifications of a function; this series sums to:

$$M_t = \frac{s - (n+1)s^{n+1} + ns^{n+2}}{(1-s)^2}$$

substituting and rearranging the first equation, and assuming  $(n+1)s^{n+1}$  is very small, gives:

$$\frac{B}{I} > \frac{(1-s)^2}{s}$$

What range of values of  $s$  occur in practice? A study by Robles, Herráiz, German and Izquierdo-Cortázar<sup>1168</sup> analysed the change history of functions in Evolution (114,485 changes to functions over 10 years) and Apache (14,072 changes over 12 years).

Figure 4.54 shows the number of functions (in Evolution) that have been modified a given number of times (upper) and the number of functions modified by a given number of different authors (lower); red line is a bi-exponential fit, green/blue lines are the individual exponentials. Data from Robles et al.<sup>1168</sup> code

The two exponentials, for number of changes gives  $s = 0.48$  ( $\frac{B}{I} > 0.58$  for Evolution, 0.47 for Apache) and  $s = 0.85$  ( $\frac{B}{I} > 0.03$ , 0.02 for Apache), the number of author fit gives  $s = 0.30$  ( $\frac{B}{I} > 1.6$  for Evolution, 3.1 for Apache) and  $s = 0.61$  ( $\frac{B}{I} > 0.24$ , 0.15 for Apache).

Is there less benefit in making an investment for the original author, do the benefits come from reducing the effort of subsequent authors in understanding the code?

The model does not include the benefit received when developers read the code without modifying it, assuming the investment was in ease of comprehension...

Figure 4.55 shows the number of modifications of a function, broken down by number of authors. this is multi-state censored data, need a fancier model... TODO

Functions are created, modified zero or more times and may be deleted before the complete package/module that contains them is deleted, or the application itself ceases to be used (??? in Evolution, ??? in Apache; many function definitions are never modified ??? in Evolution, ??? in Apache)... .

A study by Wang<sup>1423</sup> looked at the survival of clones (a duplicate sequence of 50 or more tokens) in the Linux high/medium/low level SCSI subsystems (the architecture of this system happened to have three levels). Figure 4.57 shows the survival curves, which have an initial half-life of around 18 months, but the survival rate increases over time. A proxy for code lifetime...

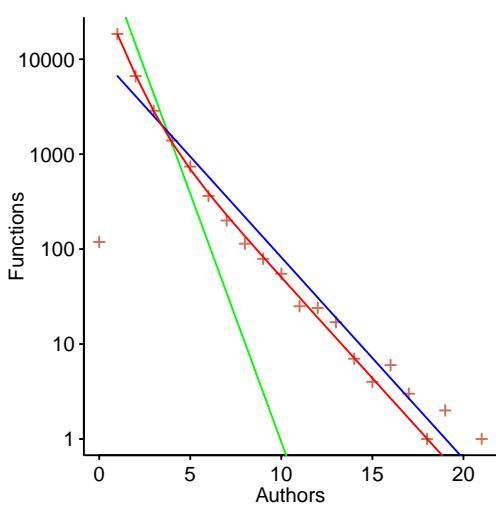
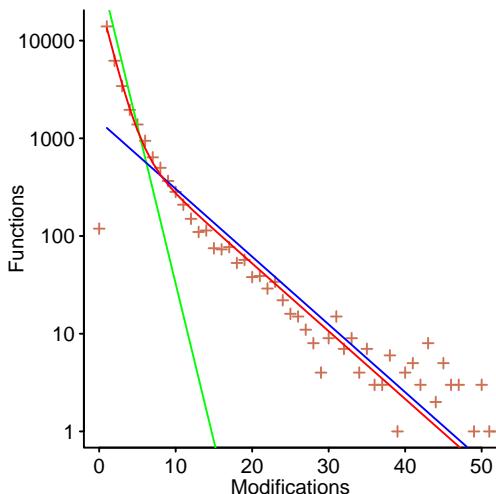


Figure 4.54: Number of functions (in Evolution; the point at zero are incorrect counts) modified a given number of times (upper) or modified by a given number of different people (lower); red line is a bi-exponential fit, green/blue lines are the individual exponentials. Data from Robles et al.<sup>1168</sup> code

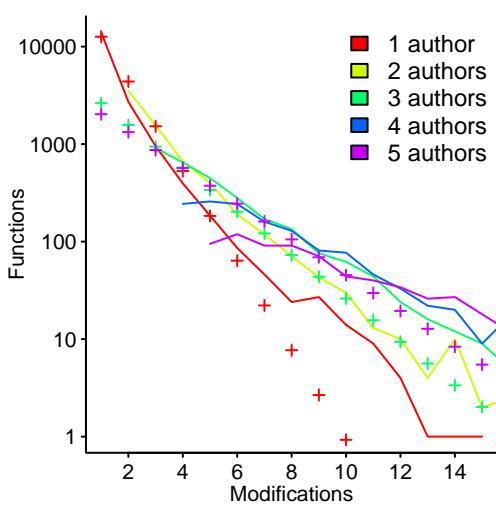


Figure 4.55: Number of functions (in Evolution) modified a given number of times broken down by number of authors. Data from Robles et al.<sup>1168</sup> code

## 4.8.2 Refactoring

Refactoring is an investment activity that assumes there will be a need to modify the code again in the future and it is more cost effective to restructure the code now, rather than at some future date. Possible reasons for time shifting an investment in code include developers not having alternative work to do or an expectation that the unknown future modifications will need to be made quickly and it is worth investing time now to reduce future development schedules; also, developers may feel peer pressure to produce code that follows accepted ecosystem norms (e.g., open source code that is about to be released)...

A study by Kawrykow and Robillard<sup>737</sup> of 24,000 change sets from seven long-lived Java programs found that between 3% and 16% of all method updates consisted entirely of non-essential modifications, e.g., renaming of local variables and trivial keyword modifications.

A study by Eshkevari, Arnaoudova, Di Penta, Oliveto, Guénéuc and Antoniol<sup>397</sup> of identifier renamings in Eclipse-JDT and Tomcat found that almost half of were applied to method names, a quarter to field names and most of the remaining to local variables and parameter names. No common patterns of grammatical form, of the renaming, were found (e.g., changing from a noun to a verb occurred in under 1% of cases). Figure 4.58 shows the number of identifiers renamed in each month, along with release dates; no correlation appears to exist between the number of identifiers renamed and releases.

## 4.8.3 Software reuse

Reusing existing code has the potential to save time and money, and reused code may contain fewer faults than newly written code (i.e., if the reused code has been executed there has been an opportunity for faults to be encountered and fixed). A surprisingly large number of equations detailing the costs and benefits of software reuse have been published;<sup>946</sup> what they all have in common is not being validated against any empirical evidence.

Multiple copies of lexically, semantically, or functionally similar source code may be referred to as *reused code*, *duplicate code* or as a *clone*.<sup>xix</sup>

Unless those making the investment needed to create reusable software receive a worthwhile benefit from its reuse, there is no incentive for making the investment. In a large organization reuse may be worthwhile at the corporate level, however the costs and benefits may be dispersed over many groups who have no incentive for investing in the larger picture.<sup>430</sup> The economics of reusable software only appears to make sense at a personal consumption level.

Reuse first requires locating code capable of being cost effectively reused to implement a given requirement. Developers are likely to be familiar with their own code and the code they regularly encounter.

A study by Li, Lu, Myagmar and Zhou<sup>841</sup> investigated copy-and-paste source within and between the subsystems of Linux and FreeBSD. Table 4.2 shows that a significant percentage of the source code of each Linux subsystem consists of replicated sequences of code; replication between subsystems is less common (the same pattern was seen in FreeBSD 5.2.1)...

subsystem	arch	fs	kernel	mm	net	sound	drivers	crypto
arc	25.1	1.4	0.5	0.3	1.1	1.3	3.2	0.1
fs	1.4	16.5	0.6	0.5	1.7	1.2	2.2	0.05
kernel	3.0	1.8	7.9	0.6	2.3	1.6	2.8	0.1
mm	2.6	2.2	0.8	6.2	1.7	1.1	2.0	0.0
net	1.8	2.5	1.1	0.7	20.7	2.1	3.7	0.1
sound	2.3	2.0	1.0	0.6	2.2	27.4	4.6	0.2
drivers	2.3	1.7	0.6	0.4	1.8	2.0	21.4	1.1
crypto	2.3	2.2	0.3	0.1	1.1	1.5	2.5	0.2
others	3.8	1.9	0.8	0.4	1.7	1.5	2.6	0.3

Table 4.2: Percentage of a module's source code cloned within and across subsystems of Linux 2.6.6. Data from Li et al.<sup>841</sup>

Are clones consistently propagated during maintenance<sup>1355</sup> ... have data...

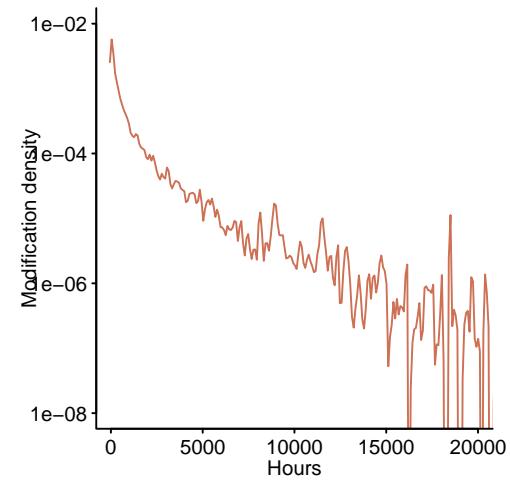


Figure 4.56: Density plot of time interval, in hours, between each modification of a function in Evolution. Data from Robles et al.<sup>1168</sup> code

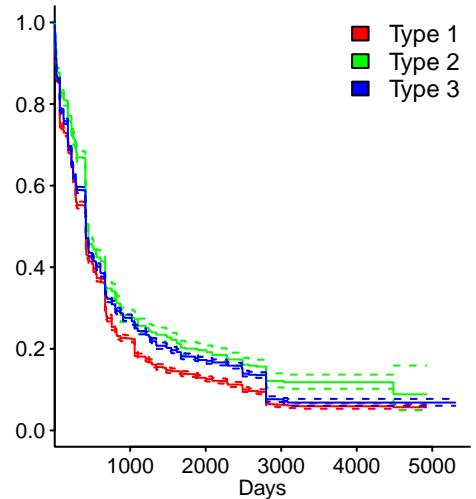


Figure 4.57: Survival curves of clones in the Linux high/medium/low level SCSI subsystems. Data from Wang.<sup>1423</sup> code

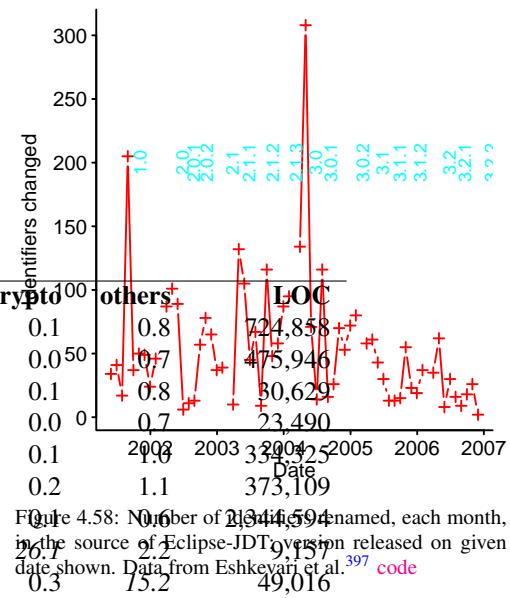


Figure 4.58: Number of renamed identifiers each month, in the source of Eclipse-JDT, version released on given date shown. Data from Eshkevari et al.<sup>397</sup> code

<sup>xix</sup> Clone is a term commonly used in academic papers.

Source code is written to implement some functionality, it either has to be written or adapted from code that found elsewhere. There have been a series of papers investigating whether cloned code is more fault prone than non-cloned code,<sup>640</sup> in general these have failed to fully control for faults in non-cloned versions of the code. There are specific faults patterns that are the result of copy-and-paste errors<sup>122</sup> and these are discussed in Chapter 6.

emailed twice (no reply...)?

Reasons for not reusing code include the cost of performing due diligence to ensure that property rights are respected (clones of code appearing on Stack Overflow<sup>xx</sup> have been found in Android Apps<sup>34</sup> having incompatible licenses), ego (e.g., being recognized as the author of functionality) and hedonism (people enjoy inventing their own wheel and will argue against using somebody else's).

#### 4.8.4 Database schema

Many applications depend on information contained in databases, with access requests often taking the form of SQL queries. The structure of a database, its schema, may change over time, e.g., columns are added/removed from tables and tables are added/removed; changes may be needed to support new functionality or a reorganization (to save space or improve performance). Changes to a schema may entail changes to the SQL embedded in application code that accesses the database.

A database's contents may be used by many applications and a change to the schema that impacts the behavior of existing applications may be costly. The cost may be high enough to outweigh the benefits of a schema change (assuming it is possible to know in advance which applications are likely to be affected), making schema changes that could impact existing code unlikely.

A study by Skoulis<sup>1264</sup> investigated changes to the database schema of various projects over time; projects included Mediawiki (the software behind Wikipedia and other wikis) and Ensembl (a scientific project involving the European Bioinformatics Institute and the Wellcome Trust Sanger Institute). Figure 4.59 suggests that table growth is approximately linear in one case (just like source code growth currently encountered in many systems, e.g., Figure 11.2), and appears to have reached its maximum in another (e.g., Figure 11.50). Systems grow in response to customer requirements...

Figure 4.60 shows the table survival curve for the Mediawiki and Ensembl database schema. Why is the table survival rate for Wikimedia much higher than Ensembl? Perhaps there are more applications making use of the contents of the Wikimedia schema, and the maintainers of the schema don't want to generate discontent in their user-base, or the maintainers are just being overly conservative. Alternatively uncertainty over what data might be of interest in the Ensembl scientific project may result in the creation of tables that eventually turn out to be unnecessary and with only two institutions involved, table removal may be an easier decision. The only way of finding out what customer demands are driving the changes is to talk to those involved.

The presence of tables and columns in a schema does not mean they are used by applications; the database administrator may be unaware that they are unused, or they may have been inserted for use by yet to be written code.

A database may contain multiple tables, with information in different tables tied together using foreign keys. A study<sup>1396</sup> of the evolution of a six databases found wide variation in the handling of foreign keys.

Any changes in newly created tables over time? Number of tables of a given size created... changes in definition of existing columns...

In some applications the database schema are contained within string literals of languages such as PHP and Python,<sup>851</sup> which makes it more difficult to analyse them...

<sup>xx</sup> Example code on Stack Overflow is governed by a Creative Commons Attribute-ShareAlike 3.0 Unported license.

## 4.9 Population dynamics

Population dynamics has been extensively studied in ecology,<sup>525</sup> which is interested in estimating species richness and diversity within a geographic region.

Individuals make choices based on their desires and goals, and within an ecosystem's population some patterns will emerge that influence subsequent individual choices. Techniques used to model population evolution include:

- mathematics, in particular differential equations: equations describing the behavior of the important variables are created and then solved; analytic solutions...

The evolution of the population of two distinct entities within a self-contained ecosystem has been studied in detail. If, say, entities  $A$  and  $B$  have fitness  $a$  and  $b$  respectively, both have a growth rate  $c$  and an average fitness of  $\phi$ , then the differential equation describing the evolution of their population size,  $x$  and  $y$ , over time is given by:<sup>1020</sup>

$$\dot{x} = ax^c - \phi x$$

$$\dot{y} = by^c - \phi y$$

Solving these equations shows that, when  $c < 1$ , both  $A$  and  $B$  can coexist, when  $c = 1$ , the entity with the higher fitness can invade and dominate an ecosystem (i.e., lower fitness eventually dies out), but when  $c > 1$ , an entity with high fitness cannot successfully invade an occupied ecosystem (i.e., greater fitness is not enough to displace an incumbent).

- simulation: covered in section...

The mathematics approach has the big advantage that the behavior of a system can be read from the equations that describe it; equations provide proofs, while simulations provide a collection of examples. The big disadvantage of the mathematical approach is that it can be very difficult to solve the equations describing many real world problems.

The advantage of simulations is that they can handle most real world problems. The disadvantages of simulations include: difficulty of exploring the sensitivity of the results to changes to model parameters, computational cost, if  $10^4$  combinations of different model parameter values are needed to cover the possible behaviors in sufficient detail, with  $10^3$  simulations for each combination (needed to reliably estimate the mean outcome), then  $10^7$  simulation runs are needed, which at 1 second each is 116 cpu days... Simulation models can be hard to communicate to others...

Many software ecosystems experience network effects<sup>56</sup>... It is in vendors' commercial interest to create what is known as a *virtuous circle*, encouraging third-party developers to sell their products within the ecosystem attracts more customers, which in turn attracts more developers and so do.

Given two new technologies, say A and B, competing for customers in an existing market, what are the conditions under which one technology becomes the market leader?

Assume that at some random time a customer has to make a decision to replace their existing technology and there are two kinds of customer: R-agents perceives a greater benefit in using technology A (i.e.,  $a_R > b_R$ ) and S-agents perceives a greater benefit in using technology B (i.e.,  $a_S < b_S$ ); both technologies are subject to networking effects; having other people using the same technology provides a benefit to everybody else using it.

Table 4.3 shows the total benefit available to each kind of customer from adopting one of the technologies;  $n_A$  and  $n_B$  are the number of users of A and B when a customer makes a decision,  $r$  and  $s$  are the benefits accrued to the respective agents from existing users (there are increasing returns when:  $r > 0$  and  $s > 0$ , decreasing returns when:  $r < 0$  and  $s < 0$ , and no existing user effect when:  $r = 0$  and  $s = 0$ ).

	Technology A	Technology B
R-agent	$a_R + rn_A$	$b_R + rn_B$
S-agent	$a_S + sn_A$	$b_S + sn_B$

Table 4.3: Returns from choosing A or B, given previous technology adoptions by others.

For increasing returns, lock-in of technology A occurs (i.e., it provides the greater benefit for all future customers) when:

$$n_A(t) - n_B(t) > \frac{b_S - a_S}{s}$$

where:  $n_A(t)$  and  $n_B(t)$  are the time dependent values of  $n$ .

The condition for lock-in of technology B is:  $n_B(t) - n_A(t) > \frac{a_R - b_R}{r}$

Starting from a market with both technologies having the same number of customers, the probability that technology A eventually dominates is:  $\frac{s(a_R - b_R)}{s(a_R - b_R) + r(b_S - a_S)}$  and technology B dominates with probability:  $\frac{r(b_S - a_S)}{s(a_R - b_R) + r(b_S - a_S)}$

For decreasing returns, both technologies can coexist.

Governments have passed laws intended to ensure that the competitive process works as intended within commercially important ecosystems (in the US this is known as *antitrust law*, while elsewhere the term *competition law* is often used). In the US antitrust legal thinking in the 1960s was based on a market structure-based understanding of competition (i.e., courts blocked company mergers that they thought would lead to anticompetitive market structures). This shifted in the 1980s, with competition assessment based on the short-term interests of consumers (i.e., low consumer prices), not based on producers or the health of the market as a whole.<sup>748</sup>

The legal decisions and rules around ensuring that the competitive process operates in the commercial market for information are new and evolving.<sup>1062</sup>

In the UK and US it is not illegal for a company to have monopoly power within a market. It is abuse of a dominant market position that gets the attention of authorities; governments sometimes ask the courts to block a merger because they believe it would significantly reduce competition.<sup>1217</sup>

Diversity and richness measures from ecology...?

#### 4.9.1 Estimating population size

It may be impossible or too costly to obtain information on all members of a population. Depending on the characteristics of the population, it may be possible to make an estimate of its size from a sample.

One technique for estimating the number of fish in a lake is to capture fish for a fixed amount of time, count and tag them before returning the fish to the lake. After allowing the captured fish to disperse the process is repeated, this time counting the number of tagged fish and those captured for the first time (i.e., untagged).

The *Chapman estimator* is an unbiased and more accurate estimate,<sup>1192</sup> than the simpler formula usually derived, i.e.,  $N = \frac{C_1 C_2}{C_{12}}$ . Assuming that all fish have the same probability of being caught and the probability of catching a fish is independent of catching any other fish:

$$N = \frac{(C_1 + 1)(C_2 + 1)}{C_{12} + 1} - 1$$

where:  $N$  is the number of fish in the lake,  $C_1$  the number of fish caught on the first attempt,  $C_2$  the number caught on the second attempt, and  $C_{12}$  the number caught on both attempts.

Applying this formula to code reviews assumes that those involved are equally skilled at finding problems, they invest the same amount of effort in the review process and all problems are equally likely to be found. When results from people of varying skill or resource availability are used, or some problems are easier to locate than others, the analysis is more complicated.

The Rcapture package supports the analysis of capture-recapture measurements where capture probability varies across items of interest and those doing the capturing. The VGAM package...

When sampling from a population whose members have been categorized in some way (e.g., by species), two common kinds of sampling data are: *abundance data* which contains the number of individuals within each species in the sample, and *incidence data* which contains a yes/no for the presence/absence of each species in the sample. Techniques for estimating population size...

The SpadeR and iNEXT packages contain functions for estimating and plotting species abundance data.

The *Chao1* estimator<sup>234</sup> gives a lower bound for the total population, based on a count of each item captured; it assumes that each member of the population has its own probability of capture, that this probability is constant over all captures and the population is sampled with replacement:

$$S_{est} \geq S_{obs} + \frac{n-1}{n} \frac{f_1^2}{2f_2}$$

where:  $S_{est}$  is the estimated number of unique items,  $S_{obs}$  the observed number of unique items,  $n$  the number of items in the sample,  $f_1$  the number of items captured once and  $f_2$  the number of items captured twice.

If a population, containing  $N$  items, is sampled without replacement, the unseen item estimate added to  $S_{obs}$  becomes:  $\frac{f_1^2}{\frac{n}{n-1}2f_2 + \frac{q}{1-q}f_1}$ ,<sup>235</sup> where:  $q = \frac{n}{N}$ .

Taking into account items occurring three and four times gives an improved lower bound.<sup>253</sup>

The ChaoSpecies function in the SpadeR package calculates species richness using a variety of models.

The number of additional items that need to be sampled,  $m_g$ , to be likely to encounter the fraction  $g$  of all the expected unique items in the population is:<sup>235</sup>

$$m_g \approx \frac{nf_1}{2f_2} \log \left[ \frac{f_0}{(1-g)S_{est}} \right]$$

where:  $n$  is the number of items in the current sample and  $f_0 = \frac{f_1^2}{2f_2}$ . For  $g = 1$ , the following relationship needs to be solved for  $m$ :  $2f_1 \left( 1 + \frac{m}{n} \right) < e^{\frac{m}{n} \frac{2f_2}{f_1}}$

If  $m$  additional items are sampled, the expected number of unique items encountered is<sup>1239</sup> (one study,<sup>277</sup> using simulation, found that reasonable estimates were given when  $m \leq n$ ):

$$S(n+m) = S_{obs} + f_0 \left[ 1 - \left( 1 - \frac{f_1}{nf_0 + f_1} \right)^m \right]$$

If  $m$  is much less than  $n$ , this equation approximates to:  $S(n+m) \approx S_{obs} + m \frac{f_1}{n}$ .

The formula to calculate the number of unique items shared by two populations is based on the same ideas, and is somewhat involved<sup>1052</sup>...

When capture probabilities vary by time and individual animal<sup>236</sup>...

?



# Chapter 5

## Projects

### 5.1 Introduction

The hardest part of any project is, generally, obtaining the funding to implement it.<sup>1216</sup>

Clients<sup>i</sup> are paying for a solution to a problem and have decided that a software system is likely to provide a cost effective solution. The client might be a large organization contracting a third-party to develop a new system, a company writing software for internal use, or an individual spending their own time implementing an idea they have.

Successfully implementing a software system involves creating a financially viable implementation that solves the client's problem. Financial viability means not so expensive the client is unable to pay for it and not so cheap that the vendor fails to make a reasonable profit.

A commercial software project aims to implement the clients' understanding of the world (in which they operate, or want to operate), in a computer executable model that can be integrated into the business and, economically operated.

Useful programs vary in size from a few lines of code to millions of lines and might be written by an individual in under an hour or by thousands of developers over many years. Much of the existing research has focused on large projects, reasons for this include: the bureaucracy needed to support large projects creates paperwork from which data can be extracted for research, and the organizations providing the large sums needed to finance large projects are able to influence research agendas. This book is driven by the availability of data and much of this data comes from large projects; small projects are important and because they are much more common than large projects, it is possible they are economically more important.

What are the characteristics of the majority of software projects? Figure 5.1 suggests that most are completed in under a year, contain less than 40 KSLOC and that much of the effort is performed by external contractors (data from a multi-year data collection process<sup>17</sup> by the Software Engineering Center, of Japan's Information-Technology Promotion Agency).

"The first go-around at it was about \$750 million, so you figure that's not a bad cost overrun for an IT project. Then I said, "Well, now, tell me. Did you do an NPV? Did you do a ROI? What did you do on a \$750 million IT investment?" And she sort of looked a little chagrined and she said, "Well, actually, there was no analysis done on that." I said, "Excuse me . . . can you explain that to me please. That's not what the textbook says." She said, "Well, it was a sales organization, the brokers worked for the sales organization." The sales organization —this was a few years ago when the brokerage business was extremely good— said, "you know, the last two years we've made more than enough money to pay for this. We want it, and we're going to pay for it." And the board of directors looked at how much money they were making and they said, "You go pay for it". So that was the investment analysis for a \$750 million IT investment that turned into a billion dollars."<sup>1335</sup>

It is unwise to ask clients why they want the software. Be thankful that somebody is willing to pay to have bespoke software written, creating employment for software developers.

Shrink-wrapped software products have a recommended retail price; finding out how much the client is willing to pay is an important part of the bespoke software sales process. James

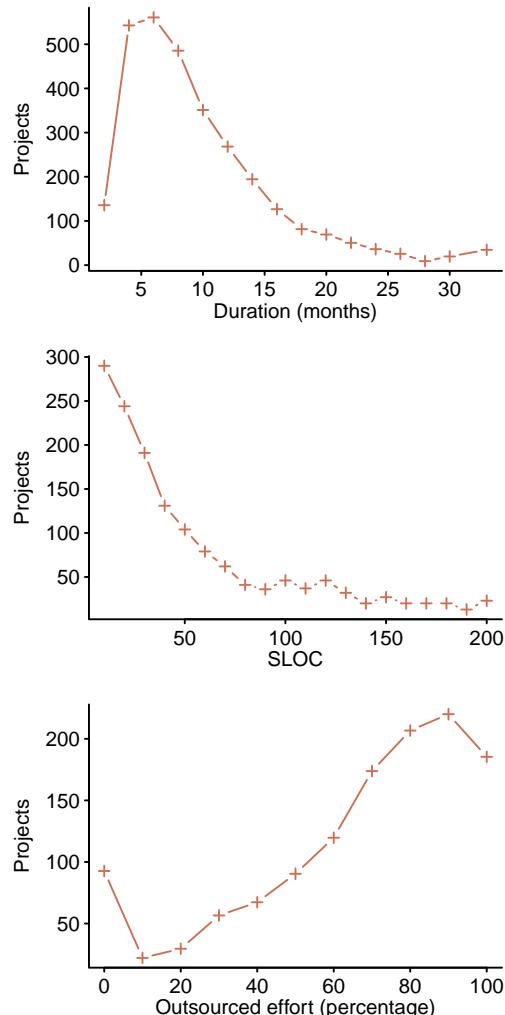


Figure 5.1: Number of projects having a given duration (upper; 2,992 projects), delivered containing a given number of SLOC (middle; 1,859 projects), and using a given percentage of out-sourced effort (lower; 1,267 projects). Data extracted from Akita et al.<sup>17</sup> code

<sup>i</sup> The term *customer* has mass market associations, bankrolling bespoke software development deserves something having upmarket connotations.

Webb, then head of NASA, told President Kennedy that \$20 billion was the price tag for landing on the Moon (NASA engineers had previously given Webb an estimated cost of \$10-12 billion,<sup>1221</sup> the final cost was \$24 billion).

The first important question a vendor needs to answer, when asked by a potential client about creating a bespoke software system, is: does the client have the money needed to pay for such a project? If the client appears to have the money and is willing to spend it, the vendor may consider it worth investing, to get a good-enough view on whether what the client thinks they want can be implemented well enough for them to pay for an implementation.

A study by Anda, Sjøberg and Mockus<sup>37</sup> sent a request to tender to 81 software consultancy companies in Norway, 46 did not respond with bids. Figure 5.2 shows the estimated schedule days and bid price received from 14 companies (21 companies provided a bid price without a schedule estimate). Fitting a regression model (making use of information on an assessment of the analysis and design performed by each company, along with estimated planned effort) explains almost 90% of the variation in the schedule estimates; see [projects/effort-bidprice.R](#).

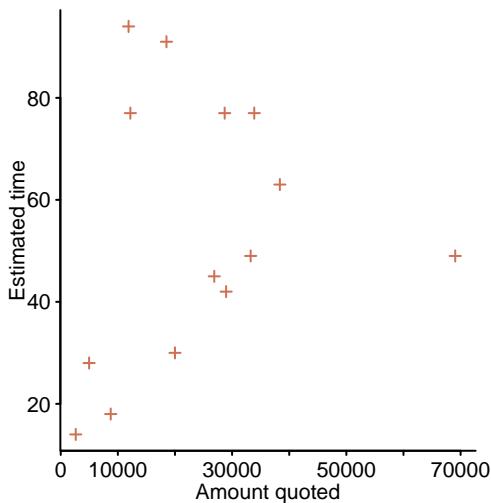


Figure 5.2: Firm bid price against schedule estimate, received from 14 companies, for the same tender specification. Data from Anda et al.<sup>37</sup> [code](#)

Projects live for as long as they are funded and can only exceed their original budget when the client is willing to pay. A project that misses its delivery date only continues to live because the client continues to agree to pay for delivery.

For some projects, clients have no choice but to pay what it takes and wait for delivery,<sup>926</sup> since not having a working software system is likely to result in the client ceasing to be competitive, resulting in ruin or a loss significantly greater than the cost of paying for the software. The funding for some projects is driven by company politics<sup>130</sup> or senior management ego, and a project might be cancelled for reasons associated with how it came to be funding, independently of any cost/schedule issues.

Both the client and the vendor want their project to be a success. What makes a project a success?

From the vendor perspective (i.e., this book's point of view) a successful project is likely to be one that produces an acceptable profit<sup>ii</sup>. A project may be a success from the vendor's perspective and be regarded as a failure by the client (because the client paid, but did not use the completed system,<sup>15</sup> or because users under-utilised the delivered system or avoided using it<sup>820</sup>) or by the developers who worked on the project (because they created a workable system despite schedules and costing underestimates by management<sup>854</sup>). These different points of view mean that results of project success surveys<sup>32</sup> are open to multiple interpretations.

A study by Milis<sup>947</sup> investigated views of project success by professionals in the roles of management, team member (either no subsequent project involvement after handover, or likely to receive project benefits after returning to their department), and end-user. Fitting regression models to the data from 25 projects, for each role, finds one explanatory variable of project success common to all roles: user happiness with the system. Management considered being within budget as an indicator of success, while other roles were more interested in meeting the specification; see [projects/Milis-PhD.R](#) for details.

A study by Garman<sup>478</sup> surveyed 70 program and project managers for US Air Force projects about their priorities. Meeting expectations (according to technical specifications) was consistently prioritized higher than being on budget or on time; none of the available explanatory variables could be used to predict priority decisions; see [projects/ADA415138.R](#) for details.

A project may fail because the users of a system resist its introduction into their workflow<sup>815</sup> (e.g., they perceive its use as a threat to their authority). Failure to deliver a system can result in the vendor having to refund any money paid by the client and make a payment for work performed by the client.<sup>61</sup>

A study by Zwikaal and Globerson<sup>1504</sup> of project cost and schedule overruns, and success in various industries, including software, suggests that except for the construction industry, overrun rates are broadly similar.

Those involved in software projects, like all other human activities, sometimes engage in subversion and lying,<sup>1178</sup> activities range from departmental infighting to individual rivalry, and motivations include: egotism, defending one's position, revenge, or a disgruntled employee.

<sup>ii</sup> Research papers often use keeping within budget and schedule as measures of success; this begs the question of which budget and which schedule, e.g., the initial, final, or intermediate versions?

### 5.1.1 Project pecking-order

Software projects are implemented within an existing cultural framework, which assigns some level of power and authority to those involved. A project's level of power and authority decides the extent to which outsiders have to adapt to the needs of the project or the project has to be adapted to the ecosystem in which it is being created. The following are some common development culture frameworks:

- one-off projects within a company which treats software as a necessary investment, which has to be made to keep some aspect of the business functioning. Company employees involved in the development might treat working on the project as a temporary secondment that is part of what they perceive to be their main job, perhaps a stepping stone to a promotion or a chance to get hands-on experience developing software (even with a view to doing this full time). External contractors may be hired to work on the project for some fixed period,
- projects within a company which derives most of its income from software products, e.g., software developed to be sold as a product. In the case of startups the founding project initiates the culture, Conway's law of organization...
- projects within a company where software is the enabler of the products and services which produce the income, e.g., embedded systems.

A study by Powell<sup>[1111](#)</sup> investigated software projects for engine control systems at Rolls-Royce. Figure 5.3 shows effort distribution (in person hours) over four projects, plus non-project work (blue) and holidays (purple'ish, at the top), over 20 months. Staff turnover and use of overtime during critical periods, means that total effort changes over time (also see Figure 11.68).

In its 2015 financial year Mozilla, a non-profit company that develops and supports the Firefox browser, had income of \$421,275 million and spent \$214,187 million on software development.<sup>[1322](#)</sup> Almost all of this income came from Google, who paid to be the default search engine used by Firefox,

- a series of one-off projects paid for by another company for internal use, i.e., contract software development. Companies in the contract software development business need to keep their employees busy with fee earning work and may assign them to distribute their time across multiple projects,
- projects where the income is the enjoyment derived from working on it; the implementation is part of the developers' lifestyle.

Single person projects are likely to experience greater variation than larger projects, not because the creator is driven by hedonism, but because a fixed release criteria may not exist and other activities may cause work to be interrupted for indefinite periods of time... Data driven by one source, rather than coming from many, is likely to show greater variance...

A few single developer projects grow to include hundreds of paid and volunteer developers. The development work for projects such as Linux<sup>[287](#)</sup> and GCC is spread over multiple organizations, making it difficult to estimate the level of commercial funding.

Code commits made by volunteer developers are less likely to occur during the working week than commits made by paid developers. Figure 5.4 shows the hourly commits during a week (summed over all commits) for Linux and FreeBSD, suggesting that Linux has many more paid developers.

The software project may be producing one component of many, with it being necessary for those on the project to work with other projects to build a larger integrated system, e.g., a spacecraft.

Academic software projects might fall within any of these categories (with personal reputation being the income sought<sup>[980](#)</sup>).

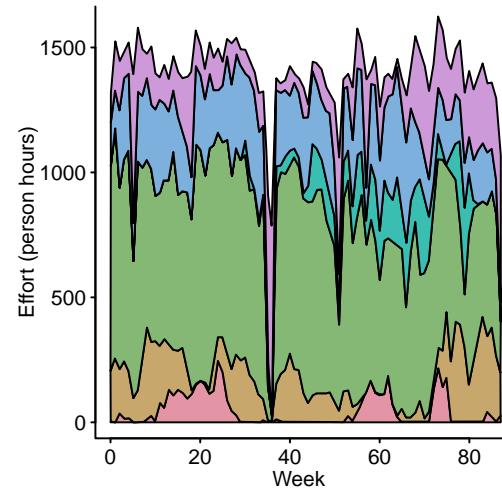


Figure 5.3: Distribution of effort (person hours) during the development of four engine control systems projects, plus non-project work and holidays, at Rolls-Royce. Data extracted from Powell.<sup>[1111](#)</sup> [code](#)

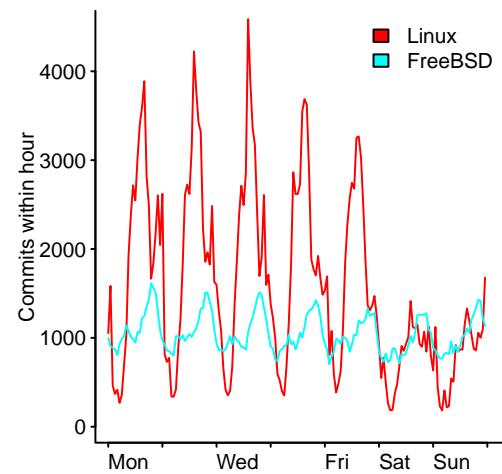


Figure 5.4: Commits within a particular hour and day of week for Linux and FreeBSD. Data from Eyolfson et al.<sup>[402](#)</sup> [code](#) July 31, 2018

### 5.1.2 Cancellation

A non-trivial percentage of projects, software and non-software,<sup>925</sup> are cancelled without ever being used. Open source and personal projects are not cancelled as-such, people simply stop working on them.<sup>269</sup> The cost-effectiveness of any investment decision (e.g., investing to create software that is cheaper to maintain) needs to include the possibility that the project may be cancelled.

- a survey by El Emam and Koru<sup>382</sup> of 84 midlevel and senior project managers, between 2005 and 2007, found the majority reporting IT project cancellation rates in the range 11-40%; Whitfield<sup>1445</sup> lists 105 outsourced UK government related ICT (Information and Communication Technology) projects between 1997 and 2007, of which 30% were terminated,
- the 1994 CHAOS report<sup>1296</sup> is a commonly cited survey of project cost overruns and cancellations. This survey is not representative of projects because it explicitly focuses on failures, subjects were asked: “ . . . to share failure stories.”, i.e., the report lists many failures and high failure rates because subjects were asked to provide this very information. The accuracy of the analysis used to calculate the summary statistics listed in the report has been questioned.<sup>401,719</sup>

Lifecycle stage	Cancelled	Completed	Overrun (schedule and/or cost)
Feasibility	0	214	0
Requirements analysis	3	211	0
Design	28	183	32
Code	15	168	57
Testing	4	164	57
Implementation	1	163	69
Handover	0	163	69
Total	23.8%	76.2%	

Table 5.1: Project cancellation rates by development stage. Data from McManus et al.<sup>921</sup>

## 5.2 Resource estimation

People tend to be overconfident and analysis shows that it pays to be overconfident; overconfidence is an evolutionary stable cognitive bias (see Section 2.8.6.2).

Agency theory deals with the conflict of interests of those paying for work to be done and those being paid; it is traditionally applied to owners of companies and those paid to manage the companies. The managers have more information than the owners and are in a position to make use of this information for personal gain, at the expense of those paying them. Developers have more information about the features they are working on and are in a position to make use of this . . .

What kind of estimate is needed, how much opportunity is there to change it later, and what are the client expectations (e.g., minimum credible and maximum willing to spend limits)?

When making a purchase decision in everyday life, people have an expectation that the seller can and will provide information on cost and delivery date, it is not considered unreasonable to expect specific answers.

Bespoke software development is not a service that many clients regularly fund; most clients want a cost and delivery date agreed before they hand over any money or sign a contract, because they perceive this to be the way things are done. Occasionally projects appear to have money thrown at them, with no fixed delivery dates attached. In these cases, the cost is likely to be very small compared to the expected benefits, but there is always a day of reckoning.

A client is paying for bespoke software because nothing suitable is available. How much will it cost to build a system that has never been implemented before, by those involved? Perhaps something very similar already exists, but unless vendors are aware of its existence

and using it has a better ROI (to them), compare to implementing something new, they have no incentive to suggest the client use it.

The following are some incentives influencing client and vendor resource estimates:

- estimating is a social process (see Figure 2.45), people often want to get along with others, which means prior estimation information can have an impact<sup>48</sup> (see Figure 5.5); client expectations have an anchoring effect on cost estimates,<sup>720</sup>?
- pressure is applied to planners<sup>1416</sup> to ensure their analysis presents a good case for project funding. A former president of the American Planning Association said:<sup>449</sup> “I believe planners and consultants in general deliberately underestimate project costs because their political bosses or clients want the projects. Sometimes, to tell the truth is to risk your job or your contracts or the next contract . . .”
- short-termism: get something out there and if enough people find it useful money for ongoing work will be found. Projects are regularly cancelled, or fail to deliver worthwhile benefits, taking short-cuts to minimise upfront costs can be a cost effective strategy...
- client unwillingness to pay for detailed analysis of proposed projects; there is no incentive for vendors to invest in detailed analysis until a contract is signed.

Believability...?

Inaccurate project resource estimates are endemic in industry<sup>936</sup> (and perhaps to all human activities), software projects are just one instance. A study by Flyvbjerg, Holm and Buhl<sup>450</sup> of 258 transportation projects (worth \$90 billion) found costs are underestimating in around 90% of projects and an average cost overrun of 28% (sd 39); a study of 35 major US DOD acquisitions<sup>155</sup> found a 60% average growth in total costs; an analysis of 12 studies,<sup>635</sup> covering over 1,000 projects, found a mean estimation error of 21%; Butts and Linton<sup>208</sup> discuss, in detail, overruns on the development of over 150 NASA spacecraft.

While the reasonableness of wanting a cost estimate cannot be disputed, an analysis of the number of unknowns involved and the experience of those involved leads to the conclusion that may be unreasonable to expect accurate resource estimates.

There is plenty of evidence<sup>iii</sup> that most resource estimates are inaccurate, with the estimate often being much lower than what is actually used. Figure 5.6 shows estimated time against actual for project sizes over four-orders of magnitude.

Cost overruns are not something the buyer or seller are likely to be want to publicise and there are probably many more instances than publicly reported.

Software project cost overruns may be blamed on poor project management,<sup>825</sup> but the poor management may have occurred during estimation. The Queensland Health Payroll System Commission of inquiry report<sup>247</sup> provides a detailed analysis of a large failed project.

?

A study by Grimstad and Jørgensen<sup>553</sup> investigated the consistency of estimates made by the same person; seven developers were asked to estimate sixty tasks, over a period of three months; unknown to them, everybody estimated (in work hours) six tasks twice. Figure 5.7 shows the first/second estimates for the same task made by the same subject; identical first/second estimates appear on the grey line and estimates for identical tasks have the same color (and extent of some color clustering shows agreement between developers).

The principles of cost estimation<sup>7</sup> are the same whatever the kind (i.e., software or otherwise) and size of project. The size of the project will have a strong influence on the amount of effort invested in forming an accurate estimate, along with the skill of those doing the estimation. For economically large projects it may be cost-effective to involve one or more expert cost estimators.

The nature of software systems development makes accurate estimation of the implementation resources needed a nonsensical goal. Duplicating everything that was done during a previous successful project implementation will deliver a project within a known cost and

<sup>iii</sup> Whitfield<sup>1445</sup> lists 105 outsourced UK public sector ICT (Information and Communication Technology) projects between 1997 and 2007, having a total value of £29.6 billion; 57% of contracts experienced cost overruns (totalling £9.0 billion) with the average cost overrun being 30.5%, and 30% of contracts were terminated.

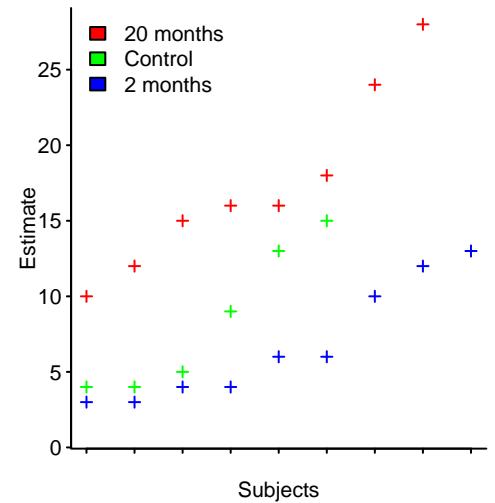


Figure 5.5: Estimate given by three groups of subjects after seeing a statement by a middle manager containing an estimate (2 months or 20 months) or no estimate (control). Data from Aranda.<sup>48</sup> [code](#)

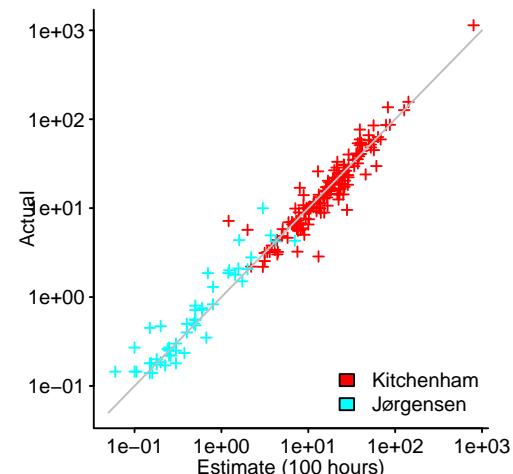


Figure 5.6: Estimated and actual project implementation effort. Data from Jørgensen<sup>710</sup> and Kitchenham et al.<sup>764</sup> [code](#)

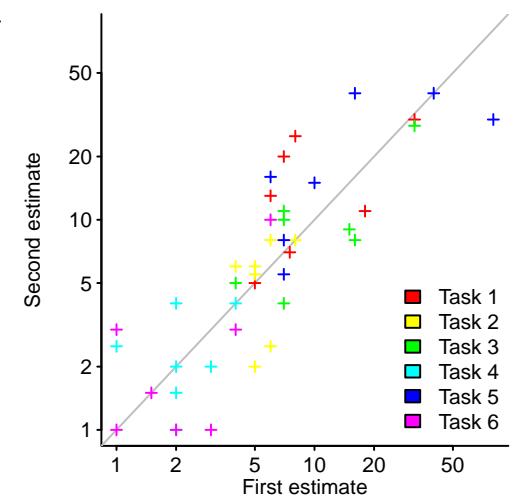


Figure 5.7: Two estimates (in work hours), made by seven subjects, for each of six tasks. Data from Grimstad et al.<sup>553</sup> [code](#)

schedule, however, duplicating everything will result in duplicate output and there are much cheaper ways of duplicating software.

A new software system is being created because no existing system is available to do the job. The reliability of any cost and time estimates depends on the accuracy with which all the necessary tasks are enumerated and the accuracy with which the resources needed to implement these tasks can be estimated.

Enumerating necessary tasks entails first enumerating client requirements (discussed below); the risk of changing requirements is a source of resource uncertainty...

Information about the resources previously used to implement the same or similar tasks is essential input to the resource estimating process for pending task; they provide a guide to the lower bounds of the resources that might be needed. When direct information is not available, indirect information is sometimes used e.g., lines of code have been used to provide a mapping to costs (see Figure 3.7).

When little, or no, information is available, following the herd can be an effective strategy.

A study by Wang and Zhang<sup>1425</sup> investigated the distribution of effort, in man-hours, used during the five major phases of 2,570 projects (the types of project were: 20% new development, 68% enhancement, 7% maintenance, 5% other projects). Figure 5.8 shows a density plot of the effort used in each phase, as a fraction of total project effort; the terminology in the legend follows that used by the authors (a mapping of those terms differing from Western usage might be: Produce is implementation, Optimize is testing and Implement is deployment). The distribution of means and medians did not vary with much with project duration.

- + Figure 5.9 shows the mean and median effort spent on projects taking a given number of months to complete (total of 2,103 projects); the lines are fitted quadratic equations. Assuming 150 man-hours per month, project team size appears to have increased from one to perhaps six or seven, as project duration increased.

Resource estimation is a knowledge based skill acquired through learning from others and practical experience (see the expertise subsection for a discussion of this issue). Making use of the prior project implementation experience of those doing the estimate is a fall-back position, when nothing better is available. It would be surprising if people were capable of estimating the resources needed to implement something that is completely new to them.

Reusing significant amounts of code complicates cost and schedule estimation.<sup>282</sup>

In some situations there is an incentive to over-estimate the time needed to complete a project. For instance, most employees are not in the position of having to undercut competitors to be given work to do. Unless the employee is under pressure, when asked by a manager, how long a task will take it may be better to over-estimate and delivery early, than to under-estimate and delivery late. The equation  $actual = estimate^{0.67}$ , explains almost half the variance in the data for Figure 8.16.

## 5.2.1 Estimation models

Estimation of computer programming costs has proved to be a complex process, with early studies<sup>411, 412</sup> identifying over fifty factors and a 1966 management cost estimation handbook<sup>999</sup> included a checklist of 94 questions (based on an analysis of 169 projects). Approaches that have been used to build software project effort estimation models include: finding equations that best fit data from earlier projects, deriving and solving theoretical models and building project development simulators.

**Fitting data:** Early cost estimation models used this approach.<sup>1441</sup> The problem with fitting equations to data, is that the resulting models are only likely to perform well when estimating projects having the same characteristics as the projects from which the data used to fit the model, was obtained. A study by Mohanty<sup>965</sup> compared the estimates produced by 12 models; Figure 5.10 shows how widely the estimates varied.

A study by Ourada<sup>1042</sup> evaluated four effort estimation models used for military software (two COCOMO derivatives REVIC and COSTMODL, plus SASET and SEER; all fitted to data); reached the conclusion: “I found the models to be highly inaccurate and very much dependent upon the interpretation of the input parameters.” Similar conclusions were reached by Kemerer<sup>741</sup> who evaluated four popular cost estimation models (SLIM, COCOMO, Function Points and ESTIMACS) on data from 15 large business data processing projects; Ferens<sup>425</sup>



Figure 5.8: Density plot of number of projects investing a given fraction of their total effort in a given project phase. Data kindly provided by Wang.<sup>1425</sup> [code](#)

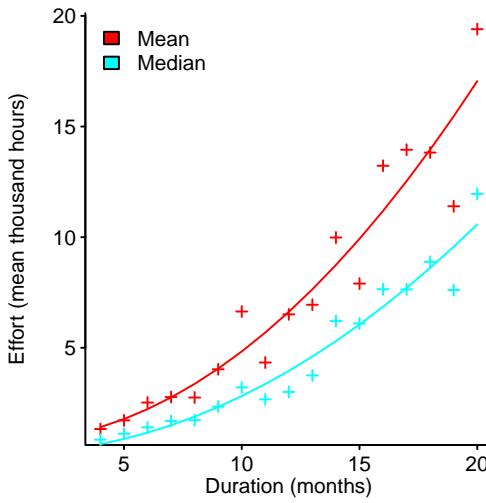


Figure 5.9: Mean and median effort (hours) for projects having elapsed time between four and 20 months (lines are fitted quadratics). Data from Wang et al.<sup>1425</sup> [code](#)

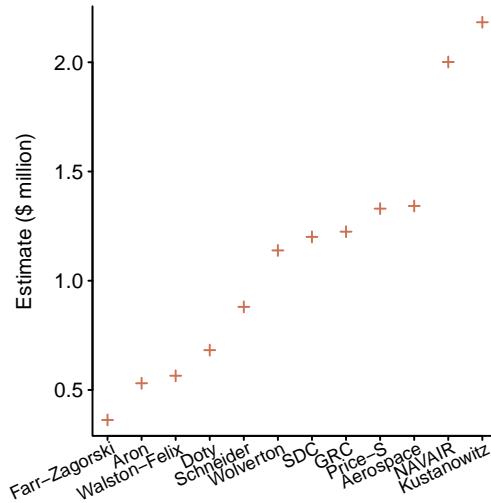


Figure 5.10: Estimated project cost from 12 estimating models. Data from Mohanty.<sup>965</sup> [code](#)

compared 10 models against DOD data and came to the same conclusion, as did another study using data from ground based systems.<sup>580</sup>

**Deriving equations:** Norden<sup>1017</sup> studied development projects, which he defined as “ . . . a finite sequence of purposeful, temporarily ordered activities, operating on a homogeneous set of problem elements, to meet a specified set of objectives . . . ”: completing a project involves solving a set of problems ( $W(t)$  is the proportion of problems solved at time  $t$ ) and these problems are solved by the people resources ( $p(t)$ , which encodes information on number of people and their skill) is ; the rate of problems solving depends on the number of people available and the number of problems remaining to be solved:

$$\frac{dW}{dt} = p(t)[1 - W(t)]$$

whose solution is:  $W(t) = 1 - e^{-\int^t p(\tau)d\tau}$

If the skill of the people resource grows linearly, as the project progresses, i.e.,  $p(t) = at$ , the work rate is:

$$\frac{dW}{dt} = ate^{-at^2/2}$$

This equation is known, from physics, as the Rayleigh curve. Putnam<sup>1125</sup> evangelised the use of Norden’s model for large software development projects. Criticism of the Norden/Putnam model has centered around the linear growth assumption (i.e.,  $p(t) = at$ ) being unrealistic.

An analysis by Parr,<sup>1058</sup> modeled the emergence of problems during a project as a binary tree and assumed that enough resources are available to complete the project in the shortest possible time. Under these assumptions the derived work rate is:

$$\frac{dW}{dt} = \frac{1}{4} \operatorname{sech}^2 \frac{\alpha t + c}{2}$$

where  $\operatorname{sech}$  is the hyperbolic secant:  $\operatorname{sech}(x) = \frac{2}{e^x + e^{-x}}$ .

While these two equations look very different, the fitted curves they produce are surprisingly similar; one difference is that the Rayleigh curve starts at zero, while the Parr curve starts at some positive value.

A study by Basili and Beane<sup>99</sup> investigated the quality of fit of various models to six projects requiring around 100 man-months of effort. Figure 5.11 shows Norden-Putnam, Parr and quadratic curves fitted to effort data for project 4.

Both the Norden/Putnam and Parr equations can be derived using hazard analysis,<sup>1400</sup> with the lifetime of problems to be discovered having a linear and logistic hazard rate respectively.

The derivation of these, or any other equation, is based on a set of assumptions, e.g., a model of problem discovery (the Norden/Putnam and Parr models assume that there are no significant changes to the requirements), and the necessary manpower can be added or removed at will. The extent to which the derived equation apply to a project depends on how closely the characteristics of the project meet the assumptions used to build the model.

**Simulation models:** A simulation model that included all the major processes involving in a development project could be used to estimate the resources likely to be needed. There have been several attempts to build such models using Systems Dynamics.<sup>2, 198, 884</sup> Building a simulation model requires understanding the behavior of all the important factors involved and as this book shows, we are a long way from having this understanding.

**Requirements based** Various techniques have been developed to estimate the resources needed to implement a set of detailed requirements. Examples include: function-points (various counting algorithms have been active used), use cases and story points.

#NoEstimates, ?

A study by Kampstra and Verhoef<sup>730</sup> investigated the reliability of function point counts. Figure 5.12 shows normalised cost for 149 projects, from one large institution, having an estimated number of function points; also see [projects/82507128-kitchenham](#).

A study by Huijgens and van Solingen<sup>648</sup> investigated two projects considered to be best-in-class, out of 345 projects, from three large organizations. Figure 5.13 shows the cost per requirement, function point and story point for these two projects over 13 releases.

A study by Commeyne, Abran and Djouab<sup>278</sup> investigated effort estimates made using COSMIC function-points and story-points. Figure 5.14 shows the estimated number of hours needed to implement 24 story points, against the corresponding estimated function-points.

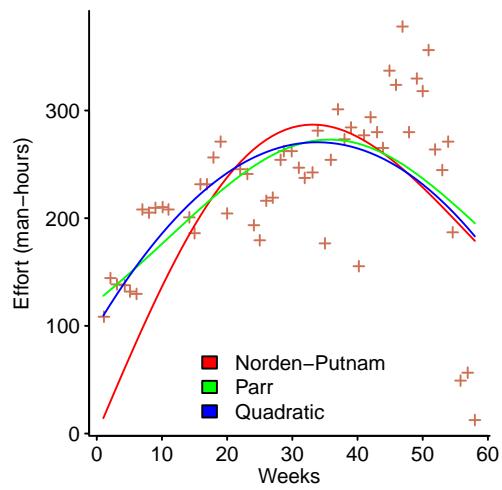


Figure 5.11: Elapsed weeks (x-axis) against effort in man-hours per week (y-axis) for a project, plus three fitted curves. Data extracted from Basili et al.<sup>99</sup> [code](#)

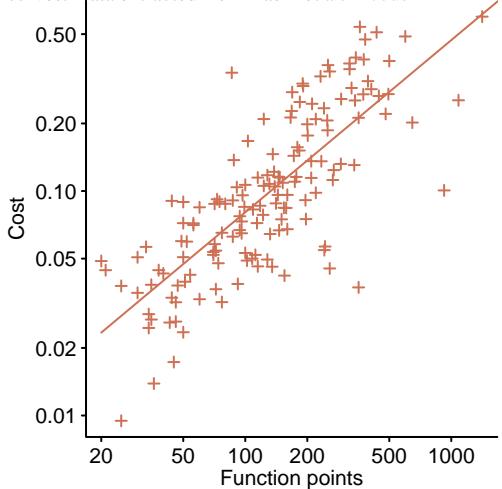
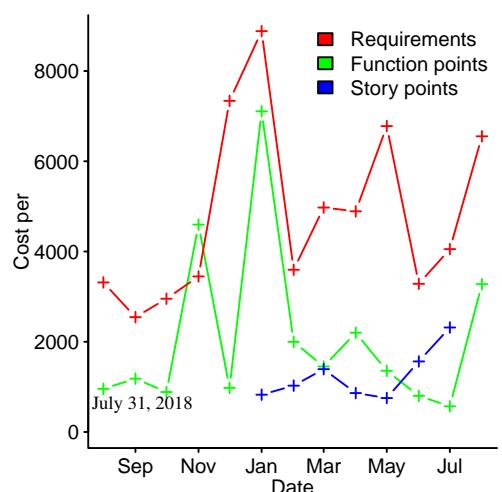
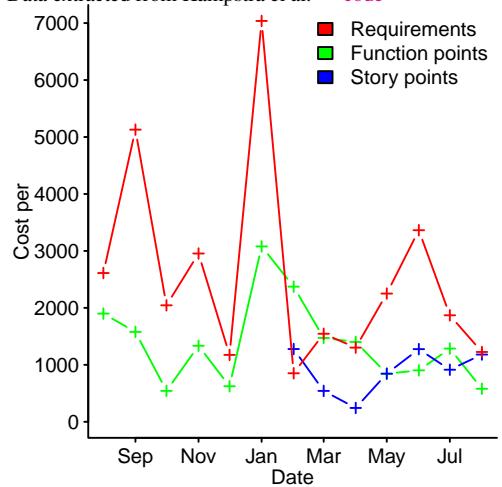


Figure 5.12: Function points and corresponding normalised costs for 149 projects from one large institution. Data extracted from Kampstra et al.<sup>730</sup> [code](#)



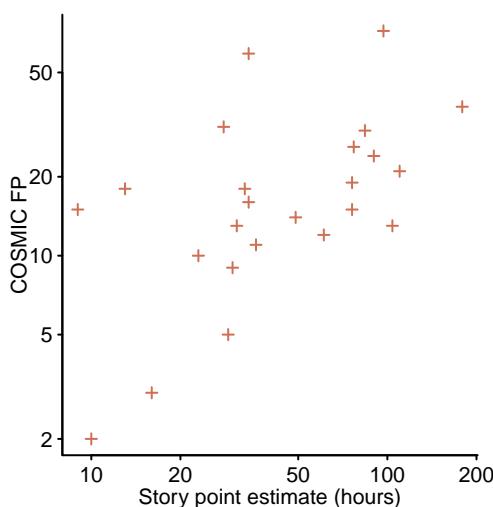


Figure 5.14: Estimated effort to implement 24 story points and corresponding COSMIC function. Data from Commeyne et al.<sup>278</sup> [code](#)

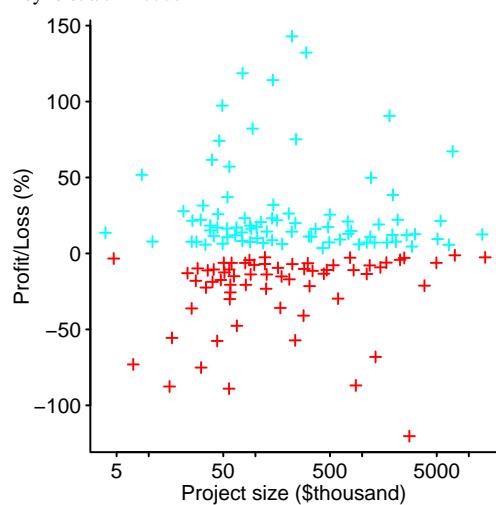


Figure 5.15: Percentage profit/loss on 145 fixed-price software development contracts. Data extracted from Coombs.<sup>285</sup> [code](#)

## 5.2.2 Money

Companies in the business of bespoke software development have to make a profit on average, over all the projects undertaken, i.e., they have some flexibility in over- and underestimating the cost of individual projects. Figure 5.15 shows the percentage loss/profit made by a software house on 146 fixed-price projects.

A study by Hill, Thomas and Allen<sup>617</sup> investigated the accuracy of over 500 task time estimates made by six managers. Around two-thirds of tasks were over estimated, one third underestimated, and 7% were on target. Over all 500+ tasks, the mean estimate was 13.51 days, actual 13.63, with standard deviations 27.83 and 27.2 respectively.

The total cost of project implementation includes the salaries of those people directly working on it, plus overheads such as office rental, support staff and equipment costs. A study by Jørgensen<sup>711</sup> selected six vendors, from 16 proposals received from a tender request, to each implement the same database-based system. The estimated price per work-hour ranged from \$9.1 to \$28.8 (mean \$13.85).

In an attempt to reduce costs some companies have offshore the development of some projects, i.e., awarded the development contract to a company in another country. A study by Šmité, Britto and van Solingen<sup>1380</sup> of outsourced project costs, found additional costs outside of the quoted hourly rates; these were attributable to working at distance, cost of transferring the work and characteristics of the offshore site. For the projects studied, likely break-even, compared to on-shore development, was thought to be several years later than planned. A study by Deutsch and Jørgensen<sup>345</sup> attempted to calculate the offshoring costs generated by what they labeled *psychic distance* (a combination of differences including cultural, language, political, geographic, and economic development).

## 5.2.3 Time

When will the system be ready for use? This is the client question that comes before or immediately after the cost request.

People time is invested to build a system; people having a minimum set of required skills and a degree of dedication. This issue is discussed in Section 5.5.

Taking longer than estimated usually means greater costs...

In some cases time and money are interchangeable on a project,<sup>839</sup> but there may be dependencies that prevent time (or money being spent) until some item is available. A cost plus contract provides an incentive to spend money, with time only being a consideration if the contract specifies penalty clauses.

Client may need to use the software by a given date...

A study by Moløkken-Østvold, Jørgensen, Tanilkan, Gallis, Lien and Hove<sup>968</sup> investigated 44 software project estimates made by 18 companies. Fitting a regression model to the data finds that estimated project duration (in elapsed days or work hours), along with kind of client (i.e., internal/external), estimation technique used and method of project organization had a statistically significant impact on actual project duration; see [projects/RK31-surveycostestim.R](#).

A study by Kitchenham, Pfleeger, McColl and Eagan<sup>764</sup> investigated the estimation accuracy of 145 projects; none of the factors measured, other than first estimate, were good predictors of actual effort; see [projects/82507128-kitchenham](#). A study by Jørgensen<sup>710</sup> investigated developer estimates for client driven product changes within a medium-sized web-development company, along with a brief description on the reason for difference between estimated/actual effort: see [projects/Regression-models.R](#).

emailed, no response... ?

## 5.2.4 Size

Program size is an important consideration when computer memory capacity is measured in kilobytes. During the 1960s mainframe computer memory was often measured in kilobytes, as was minicomputer memory in the 1970s and microcomputer memory in the 1980s. Today, low cost embedded controls might contain a kilobyte of memory, or less, e.g., electronic control units (ECU) used in car engines contain a few kilobytes of flash memory.

Figure 5.16 shows IBM's profit margin on sales of all System 360's sold in 1966 (average monthly rental cost, for 1967, in parentheses). Low-end systems, with minimal memory, were sold below cost to attract customers (and make it difficult for competitors to gain a foothold in the market<sup>339</sup>).

A study by Lind and Heldal<sup>1855</sup> investigated the possibility of using COSMIC function points to estimate the compiled size of software components used in ECUs. Function points were counted for existing software components in ECUs used by Saab and General Motors; Figure 5.17 shows COSMIC function points against compiled size of components from four ECU modules; lines are a fitted regression model for each module. While a consistent mapping exists for components within each module, the function point counting technique used did not capture enough information to produce consistent results across modules.

How much variation is to be expected in the size of programs (perhaps measured in lines) implementing the same functionality, in the same language?

Figure 9.13 shows the number of lines contained in over 6,300 C programs implementing the  $3n + 1$  problem (mean lines of code is 46). A more substantial example is the number of statements and declarations (comments or blank lines were not counted) in five Pascal compilers targeting the same mainframe.<sup>1248</sup>

Figure 5.18 shows data from seven studies where multiple implementations of the same specification were implemented in the same language. The fitted regression (grey line) finds, to a good approximation, that standard deviation is one quarter of the mean. With so much variation, in LOC, between different implementations, a wide margin of error must be assumed for any estimation technique based on lines of code.

Multiple teams developing the same project in different languages...?

Lines of code are the end result of decisions made at higher levels of abstraction, where different implementations intended to perform similar activities vary in how their internal components are structured.<sup>772</sup>

## 5.3 Pitching for projects

Bidding on tenders, convincing senior management to fund a project, selling the benefits of a bespoke solution to potential clients, all involve making commitments that directly impact project implementation. An appreciation of some of the choices that might be made while pitching for a project may be useful for understanding why things are they way they are.

A client interested in acquiring bespoke software may quickly change their mind, if they hear a development cost that is much higher than expected, or a delivery date much later than wanted. Optimal frog boiling entails starting low and increasing at a rate that does not disturb...

If it has been decided that a project will be implemented internally, within the company, then company politics decides who does what.

Companies and government departments needing bespoke software may put together a list of requirements, along perhaps with a specification, and invite interested parties to submit bids to implement the system.

Companies in the business of developing bespoke software need to maintain a pipeline of projects, opening with client qualification and closing with contract signature.<sup>1276</sup> Acquiring paying project work is the responsibility of the sales department and is outside the scope of this book.

Many of the factors involved in project bidding are likely to be common to engineering projects in general, e.g., highway procurement.<sup>87</sup> Bidding decisions can be driven by factors that have little or no connection with the technical aspects of software implementation, or its costs; some of these include:

- probability of being successful on bids for other projects. If work is currently scarce, it may be worthwhile accepting a small profit margin, or even none at all, simply to have work that keeps the business running. A software development organization, whether it contains one person or thousands, needs to schedule its activities to keep everyone busy and the money flowing in.

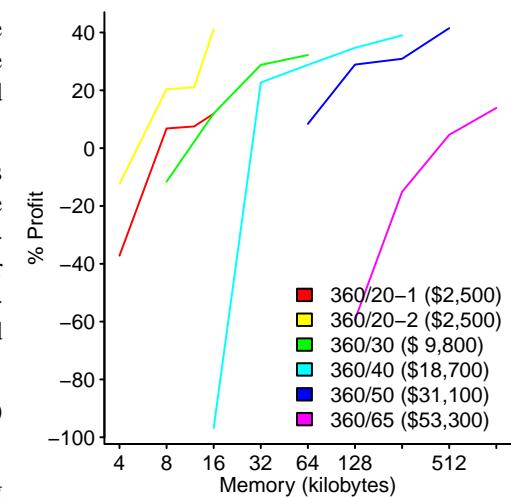


Figure 5.16: IBM's profit margin on all System 360s sold in 1966, by system memory capacity in kilobytes; monthly rental cost during 1967 in parentheses. Data from DeLaMatser.<sup>339</sup> code

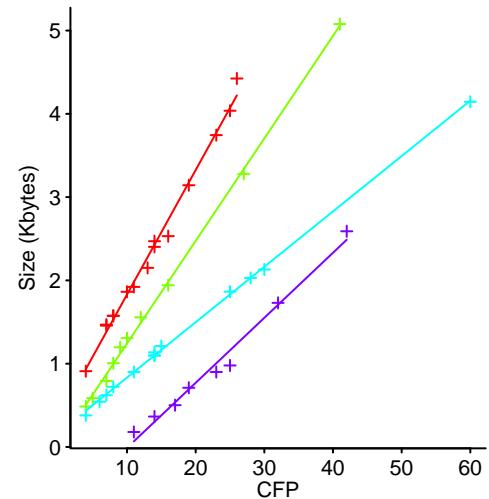


Figure 5.17: COSMIC function-points and compiled size (in kilobytes) of components in four different ECU modules; lines show fitted regression model. Data from Lind et al.<sup>1855</sup> code

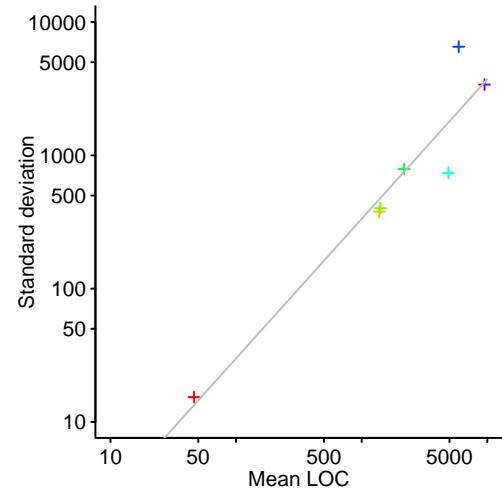


Figure 5.18: Mean LOC against standard deviation of LOC, for multiple implementations of seven distinct problems; grey line is fitted regression. Data from: Anda et al.<sup>37</sup> Jørgensen,<sup>711</sup> Lauterbach,<sup>821</sup> McAllister et al.,<sup>908</sup> Selby et al.,<sup>1225</sup> Shimasaki et al.,<sup>1248</sup> van der Meulen.<sup>1389</sup> code

Some of the schedule estimates in Figure 5.2 can be explained by companies assigning developers to more than one project at the same time, improving staff efficiency by having something else for them to do when they are blocked from working on their primary project,

- bid what the maximum price the client is currently willing to pay. A profitable strategy when the client estimate is significantly greater than the actual; if the client perceives a project to be important enough, they are likely to be willing to pay more once existing monies are spent. It is better to pay £2 million for a system that provides a good return on investment than the £1 million actually budgeted, if the lower prices system is not worth having,
- the likely value of estimates submitted by other bidders; competition is a hard task master,
- bidding low to win and ensure that wording in the contract allows for increases due to unanticipated work. Prior experience shows that clients often want to change the requirements, estimating the cost of the new requirements occurs after the competitive bidding process (see Figure 3.4). The report by the Queensland health payroll system commission of inquiry<sup>247</sup> offers some insight into this approach. Clients often prefer to continue to work with a supplier who has run into, even substantial, difficulties,<sup>162</sup>
- bidding low to win, planning to make substantial profits during the maintenance phase. This strategy is based on having a reasonable expectation that the client will use the software for many years and that the software will need substantial maintenance,
- bidding on projects that are small, relative to the size of the development company, as a means of getting a foot in the door to gain access to work involving larger projects, e.g., becoming an approved supplier.

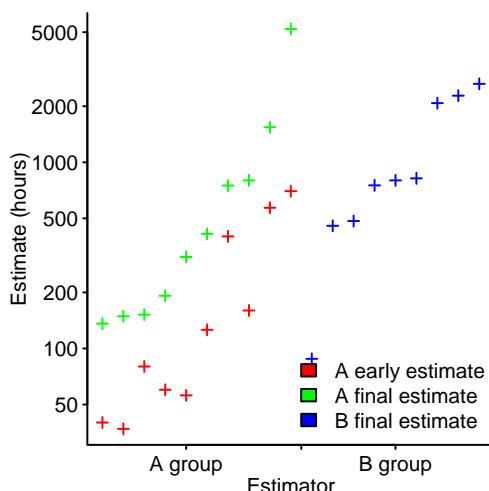


Figure 5.19: Bids made by 19 estimators from the same company (divided into two groups for the experiment). Data from Jørgensen et al.<sup>713</sup> code

A study by Jørgensen and Carelius investigated the impact of changes in the project specification on the amount bid; estimators in group A made an estimate based on a one-page specification and then a second estimate based on an eleven-page specification; estimators in group B made an estimate based on the eleven-page specification only. Figure 5.19 shows bids made by two groups of estimators from the same company; for additional analysis, see [projects/proj-bidding.R](#).

Estimates will be affected by the characteristics of cognitive processing of numeric values (see Section 2.7.1), e.g., the measurement units used (such monthly or yearly)<sup>1377</sup> can lead to different answers being given.

The bidding on a project for a new IT system for Magistrates' Courts (the Libra project),<sup>162</sup> started with ICL submitting a bid of £146 million, when it became public there was only one bidder this was increased to £184 million over 10.5 years, a contract was signed, a revised contract was renegotiated for £319 million, ICL threatened to repudiate the renegotiated contract and proposed a new price of £400 million, reduced its proposed price to £384 million and after agreement could not be reached signed a revised contract for £232 million over 8.5 years.

### 5.3.1 Contracts

A signed contract is the starting point of serious investment of resources in project implementation (some projects never involve a contract, and work may start on a project before a contract is signed). Having to read a contract after it has been signed is an indication that one of the parties involved is not happy; nobody wants to have to deal with lawyers.<sup>913</sup>

What is the payment schedule for the project? The two primary contract types are *fixed price* and *time and materials* (also known as *cost plus*; where the client pays the vendors costs plus an agreed percentage profit. e.g., a margin of 10-15%<sup>279</sup>).

On projects of any size, agreed amounts are paid at specified milestones. Milestones are a way for the client to monitor progress and the vendor to receive some income for the work they have done. The use of milestones favours a sequential development viewpoint and one study<sup>59</sup> found that the waterfall model is effectively written into contracts.

From the client's perspective a fixed price contract appears attractive, but from the vendor's perspective this type of contract may be unacceptably risky (one study<sup>519</sup> found that vendors preferred a fixed price contract when they could increase their profit margin by leveraging

particular staff expertise; another study<sup>844</sup> found that fixed-price was only used for trusted vendors). Writing a sufficiently exact specification of what the system is expected to do, along with tightly defined acceptance criteria is time-consuming and costly, which means the contract has to contain a mechanism to handle changes to the requirements; such a mechanism that is open to exploitation by both clients and vendors.

A time and materials contract has the advantage that vendors are willing to accept open-ended requirements, but has the disadvantage (to the client) that the vendor has no incentive to keep costs down.

Contracts sometimes include penalty clauses and incentive fees (which are meaningless unless linked to performance targets<sup>389</sup>).

if client cancels a project, the vendor contract should have been written to ensure that a profit is made on the work done... Depending on reason for cancellation, to what extent does reputation may suffer...

A study by Webster<sup>1436</sup> analysed legal disputes involving system failure filed in the period 1976-2000 (120 were found). The cases could be broadly divided into seven categories: client claims the installed system is defective in some way and vendor fails to repair it, installed system does not live up to the claims made by the vendor, a project starts and the date of final delivery continues to slip and remain in the future, unplanned obsolescence (client discovers that the system either no longer meet its needs or that the vendor will no longer support it), the vendor changes the functionality or configuration of the system resulting in unpleasant for one or more client, a three-way tangle between vendor, leasing company and client, miscellaneous.

Many commercial transactions are governed by standard form contracts. A study by Marotta-Wurgler<sup>897</sup> analysed 647 software license agreements from various markets; almost all had a net bias, relative to relevant default rules, in favor of the software company (who wrote the agreement).

The issue of personal integrity is discussed in Section 2.8.7.

Require rights to the source code, have to ensure that developers do not make inappropriate use of code (e.g., downloaded from the Internet<sup>1277</sup>) having incompatible licenses...

?

Client and vendor are in an asymmetric information situation... moral hazard...

?

could extract data...?

emailed, could extract... ?, ?, ?

emailed for data... ?

## 5.4 The path to delivery

A project to deliver a working system involves researching, in detail, what the client wants (i.e., requirements gathering), formulating a design<sup>303</sup> that will do this, implementing it, and testing to ensure an acceptable level of confidence in its behavior (testing is covered in Chapter 6);<sup>979</sup> for successful systems, deployment to customers marks the start of another project, ongoing maintenance and upgrades.

Managing the implementation a software system is an exercise in juggling the available resources, managing the risks (e.g., understanding what is most likely to go wrong<sup>72</sup> and having some idea about what to do about it), along with keeping the client convinced<sup>iv</sup> that a working system will be delivered within budget and schedule; these are all standard project management issues.<sup>768</sup> Managements interest in technical issues focuses on the amount of resource they are likely to consume and the developer skill-set needed to implement them.

Progress reports detailing poor progress and increased costs may be ignored,<sup>739</sup> or trigger an escalation of commitment<sup>v</sup>...

---

<sup>iv</sup> Successful politicians focus on being elected and then staying in office;<sup>322</sup> successful managers focus on getting projects and then keeping the client on-board; unconvincing clients cancel projects, or look for others to take it over.

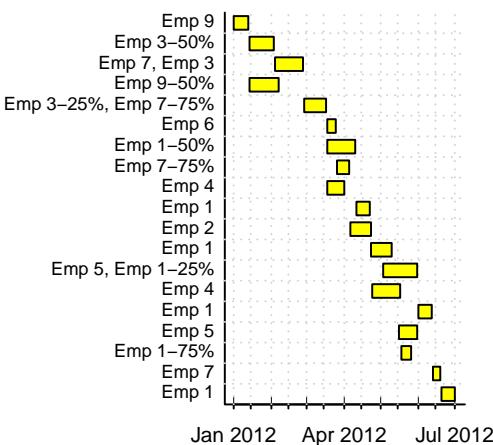


Figure 5.20: Initial implementation schedule, with employee number(s) given for each task (percentage given when not 100%) for a project. Data from Ge et al.<sup>483</sup> code

?

?

Some of the risks include the client changing the requirements <sup>v</sup>, proposed solutions, to problems that have not been solved before (at least by the people available to work on the project) not working as planned, the discovery of new problems and changes to the known problems generating costs and schedule delays, and the hardware not being capable of doing what is required...

Developers and clients can very different views about the risks involved.<sup>797</sup>

Figure 5.20 shows an implementation schedule for one of the projects studied by Ge and Xu.<sup>483</sup> This schedule is based on breaking the project into 19 components, with some component dependencies known in advance along with the skills needed to implement them, information on the expected availability of developers having given skills was enough to create a schedule; as the implementation progresses the information is uncovered may result in schedule changes, e.g., more components dependencies, and available resources changes (such as developers ceasing to be available). This plot gives the impression of a project under control, in practice slippage on another project can prevent scheduled staff being available, a requirement change can generate a new dependency that prevents a subcomponent being implemented concurrently; schedules evolve, sometimes beyond all recognition.

A study by Yu<sup>1476</sup> investigated the development of three multi-million pound projects. The chronology of events documented gives some insight into how responsibility for the implementation of functionality, and associated costs, can shift between project contractors as new issues are uncovered and contracted budgets are spent. Figure 5.21 shows the changes in contractors' project cost estimates, for their own work, over the duration of the project.

A project has a beginning, an ending and for non-trivial projects milestones in-between. Every person who works on a project has their own beginning and ending.

When does a project start? For one person, work on a project starts when they first think of bidding on a contract, writing a proposal or starting to spend time thinking about an idea. Developers might not even know about the project they are about to work on until a contract has been signed, and they are allocated to work on it.

When does a project end? Depending on the kind of project being developed, it might finish when the client makes the final contract payment, a product is first offered for sale, or the code is no longer considered to be a release candidate; once the system is being used a shift to a maintenance mode of working may feel like a continuation, under another name, for some. Cancellation is an ending that nobody wants.

Requirements, as perceived by the client, change, leading to the inevitable request for the project to support an updated list of requirements.

Vendors want to keep the client happy, but not if it means losing a significant amount of money. A good client relationship manager will know when to tell the client that extra money is needed to support the requested change and when to accept it without charging (any connection with implementation effort is only guaranteed to occur for change requests having significant cost impact)...

Project implementation strategy is strongly influenced by the cost impact of changes to requirements; the risk cost profile of changes...

When there is a very high cost associated with failing to detect an important requirement, it is worth making a significant investment in having a high degree of confidence that all the necessary requirements are known, e.g., see Figure 5.30.

Which requirements drive implementation costs and how likely are they to change? Brand<sup>172</sup> introduced the concept of *shearing layers* to refer to the way buildings contain multiple layers of change (Lim<sup>850</sup> categorised the RALIC requirements into five layers)...

A study by Simcoe<sup>1254</sup> investigated the modularity of communication protocol standards involved in supporting the Internet. Figure 5.22 shows the number of citations from IEFT and W3C Standard documents, grouped by protocol layer, that reference Standard documents in the same and other layers. Treating citations as a proxy for dependencies, 89% are along the main diagonal (a uniform distribution would produce 17%), a value that might be used to

<sup>v</sup> A well written contract includes provisions for the extra costs and schedule slippage.

estimate the extent to which dependencies will impact concurrent implementation of multiple protocols; dependencies discovered during implementation may substantially change the initial estimate.

Existing requirements may be modified on the basis of what the code does, rather than what the specification said it should do...<sup>198</sup>

Agreement between *clients* about whether requirements have been met...<sup>1115</sup>

The issue of ambiguity in requirement specifications is discussed in Chapter 6.

Gause and Weinberg<sup>481</sup> provide a readable tour through the various aspects of requirements handling in industry.

Use cases... Aimed towards specifying what a system is supposed to do... extract data from paper?<sup>9</sup>

list of features removed from Linux... [feature-removal-schedule.txt](#)

Daily client contact has been found to have an impact on final project effort, see Figure 11.42. The study did not collect any data on the extent to which the planned project goals were achieved and it is possible that daily contact enabled the contractor to convince the client to be willing to accept a deliverable that did not support the functionality originally agreed upon...<sup>1254</sup>

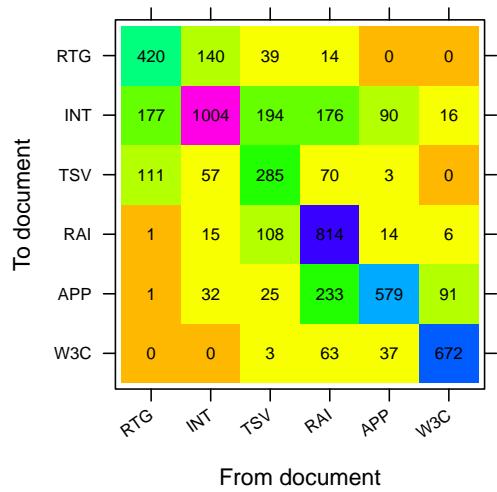


Figure 5.22: Number of citations from Standard documents within protocol level to documents in the same and other levels (RTG routing, INT internet, TSV transport, RAI realtime applications and infrastructure, APP Applications, W3C recommendations). Data from Simcoe.<sup>1254</sup>  
code

## 5.4.1 Development methodologies

The difference between projects in established engineering disciplines (e.g., bridge building in Civil engineering) and projects in disciplines where empirically established practices have not yet been established (e.g., software development) is that in the former projects are mainly a problem of how to make the best use of known alternatives, while the latter involves discovering what the workable alternatives might be (e.g., learning by trying ideas until something that works well enough is found).

The *Waterfall model* continues to haunt software project management, despite repeated exorcisms over several decades.<sup>817</sup> The paper<sup>1182</sup> that gave birth to this demon meme contained a diagram, with accompanying text claiming this approach was risky and invited failure, with the diagrams and text on subsequent pages showing the recommended approach to producing the desired outcome. The how not to do it approach, illustrated in the first diagram, was taken up, becoming known as the waterfall model and included in the first version of an influential standard, DoD-Std-2167,<sup>342</sup> as the recommended project management technique.<sup>vi</sup> Projects have been implemented using a Waterfall approach,<sup>1461</sup> but it is not considered to be the approach most likely to succeed. data...

The U.S. National Defense Authorization Act, for fiscal year 2010, specified that an iterative acquisition process be used for information technology systems; Section 804:<sup>875</sup> contains the headline requirements “(B) multiple, rapidly executed increments or releases of capability; (C) early, successive prototyping to support an evolutionary approach; . . .”. However, the wording used for the implementation of the new process specifies: “. . . well-scoped and well-defined requirements.”, i.e., bureaucratic inertia holds sway.

The battle over the communication protocols used to implement the Internet is perhaps the largest example of a waterfall (the OSI seven-layer model, ISO 7498 or ITU-T X.200, documented the requirements first, which vendors could then implement) vs. iterative approach (the IETF process was/is based on rough consensus and running code<sup>1187</sup>).

The economic environment and power structure in which a project is developed can have a significant impact on the management techniques used. For instance, the documentation and cost/schedule oversight involved in large government contracts requires a methodology capable of generating the necessary cost, schedule and documents; in winner take-all markets there is a strong incentive to be actively engaging with customers as soon as possible, and a methodology capable of producing regularly updated working systems provides a means of rapidly responding to customer demands as well as reducing the delay between writing code and generating revenue from it.

<sup>vi</sup> Subsequent versions removed this recommendation and more recent versions recommend incremental and iterative development. The primary author of DoD-Std-2167 expressed regret for creating the strict waterfall-based standard, that others advised him it was an excellent model; in hindsight, had he known about incremental and iterative development, this would have been strongly recommended, rather than the waterfall mode.<sup>817</sup>

A study by Rico,<sup>1159</sup> going back over the 50-years, identified 32 major techniques (broken down by hardware era and market conditions) that it was claimed would produce savings in development or maintenance. There have been a few field studies of the methodologies actually used (rather than what managers claim to be using); system development methodologies have been found to provide management support for necessary fictions,<sup>992</sup> e.g., a means to creating an image of control to the client or others outside the development group.

Iterative development has been independently discovered many times.<sup>817</sup> The advantages of iterative development include: not requiring a large upfront investment in requirements gathering, the potential to reduce wasted effort implementing unwanted requirements (through feedback received from users of the current system), working software makes it possible to start building a customer base (a crucial requirement in winner take-all markets) and the lag between writing code and earning income from it is reduced.

The additional cost paid for using iterative development, compared to an upfront design approach, is having to constantly reengineer code, to adapt it to support functionality that it was not originally designed to support.

If the cost of modifying the design is very high, it may be cost effective to do as much design as possible before writing code; the cost of design changes may be high when: hardware is involved, many organizations are involved and working to one design...

When customer requirements are rapidly changing, or contain many unknowns, the design may get out-of-date quickly, and it may be more cost effective to actively drive the design by ship something working (e.g., evolving series of languages and their corresponding compilers;<sup>101</sup> see projects/J04.R). User feedback is the input to each iteration, and without appropriate feedback iterative projects can fail.<sup>533</sup>

The Internet provides an ideal ecosystem for the use of iterative techniques. During the growth of Internet usage the requirements were uncertain and subject to rapid change, many market niches had winner-take-all characteristics and the Internet provided a distribution channel for frequent software updates...

The extent to which techniques used to implement previous projects are applicable to the current project... If the exact project has not been implemented before, then projects containing very similar features can be used as a guide.

Test driven development (TDD) emailed for more (available data is poor)...?

Pair programming... emailed...?

Survey data...? Bayesian Statistics in Software Engineering<sup>464</sup> ... version-one/

A study by Özbek<sup>1044</sup> investigated attempts to introduce software engineering innovations into 13 open source projects in a one-year period. Of the 83 innovations identified in the respective project email discussions 30 (36.1%) were successful, 37 (44.6%) failed and 16 (19.3%) classified as unknown.

## 5.4.2 Managing progress

How is the project progressing towards the target, in cost and time, when a project deliverable is accepted by the client?

A project involves work being done and people doing it. Work and staff have to be meshed together in a schedule; project managers will have a view on what has to be optimized. A company employing a relatively fixed number of developers may seek to schedule work so that employee always have something productive to do, while a company with a contract to deliver a system by a given date will seek to schedule staff to optimise for delivery dates.

Scheduling a software project involves solving the kinds of problems encountered on many other kinds of creation projects, e.g., staff availability (in time and skills), dependencies between work items<sup>167</sup> and other projects, and hardware resources. Software specific issues include the difficulty of finding people with the necessary skills and high turnover rate of current staff. Project staffing is discussed in Section 5.5.

Tracking progress... Earned value management...

Analyzing project progress data in isolation may be misleading. A study by Curtis, Sheppard and Kruesi<sup>304</sup> investigated the characteristics (e.g., effort and faults found) of the implementation of five relatively independent subcomponents of one system; Figure 5.23 shows how

effort, in person hours, changed as the projects progressed. While the five subcomponents were implemented by different teams, it is not known whether team members worked on other projects within their company. Figure 5.3 shows how, at a multi-project level, total available effort may not significantly change, but large changes can occur at the single project level (because higher management optimise staffing across projects).

?

While it may be obvious to those working on a project that the schedule will not be met, nobody wants to be the bearer of bad news and management rarely have anything to gain by reporting bad news earlier than they have to.

Once the client believes the estimated completion date and/or cost is not going to be met, the project is either scaled back, cancelled or a new completion estimate is accepted. Scheduling when to tell clients about project delivery slippage and/or a potential cost overrun is an essential project management skill.

The uncertainty around the schedule of a successful project is likely to decrease as it progresses towards completion. The metaphor of a *cone of uncertainty* is sometimes used for analysing project uncertainty; in practice this metaphor is at best useless. Schedule uncertainty is asymmetric, with underestimates dominating, the x-axis is calculated using a quantity that unknown until project completion (i.e., the duration of the project) and curved boundary seen in plots are a mathematical artifact. Figure 5.24 shows a plot of project percentage actually completed against the ratio  $\frac{\text{Actual}}{\text{Estimated}}$  for the corresponding project (4.6% of estimate completion dates are less than the actual date); the curved boundary does not contain any information, it is purely a mathematical artifact created by the choice of axis, i.e., the following holds:

$$\frac{\text{Actual}}{\text{Estimated}} \leq x_{\text{percentage}} \cdot \text{Actual}$$

cancelling *Actual* gives:  $\frac{1}{\text{Estimated}} \leq x_{\text{percentage}}$ ; whatever the value of *Estimated*, its point always appears below or on the  $\frac{1}{x}$  curve in the plot.

A study by Little<sup>857</sup> investigated schedule estimation for 121 projects at Landmark Graphics between 1999 and 2002; an estimated release date was made at the start and whenever the core team reached consensus on a new date (average 7.2 estimates per project). The extent to which the schedule estimation characteristics found in Landmark projects, for the period of the data, will depend on corporate culture and requirements volatility. The Landmark Graphics corporate culture viewed estimates as targets, i.e., “what’s the earliest date by which you can’t prove you won’t be finished?”,<sup>857</sup> different schedule characteristics are likely to be found in a corporate culture that punishes the bearer of bad news.

Figure 5.25 shows the number of release date estimates made for 121 projects, for a corresponding initial estimated project duration (on x-axis).

Reasons for failing to meet a project deadline include (all starting points for lawyers looking for the best strategy to use to argue their client’s case after a project fails<sup>171</sup>): an unrealistic schedule, a failure of project management, changes in the requirements, encountering unexpected problems, staffing problems (i.e., recruiting or retaining people with the required skills) and being blocked because a dependency is not available.<sup>167</sup> Missed deadlines are common and the usual response is to release an updated schedule; after too many missed deadlines projects are cancelled... data...

When is a new estimate for a project’s release date most likely to occur? If a lot of effort was invested in analyzing a project before it started, then new information, relevant to delivery date, is less likely to be discovered at the start of a project, compared to when little upfront analysis has been made. As the estimated delivery date approaches, it becomes easier for team members to judge the likelihood of delivery occurring.

If management want to continue development of a project after it has failed to be delivered by the scheduled date, they have to ask for the necessary resources; those paying for the work have to be given sufficient notice about the expected need for more resources, so that those resources can be made available (or not).

Figure 5.26 shows 882 changed delivery date announcement, with percentage elapsed time when the estimate was announced (based on the current estimated duration of the project) along the x-axis, and percentage change in project duration (for the corresponding project)

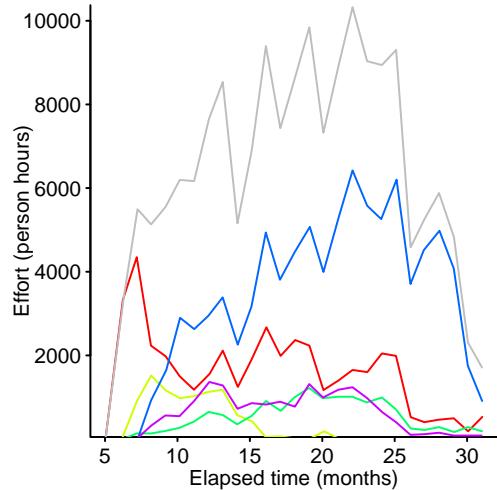


Figure 5.23: Effort, in person hours per month, used in the implementation of the five components making up the PAVE PAWS project (grey line shows total effort). Data extracted from Curtis et al.<sup>304</sup> code

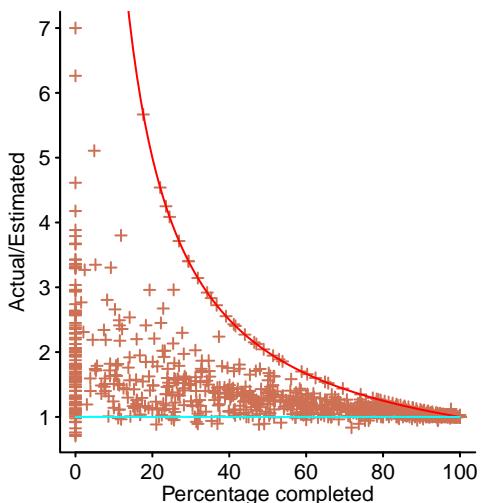


Figure 5.24: Percentage of actual project duration elapsed when 882 schedule estimates were made, during 121 projects, against estimated/actual time ratio (boundary maximum in red). Data kindly provided by Little.<sup>857</sup> code

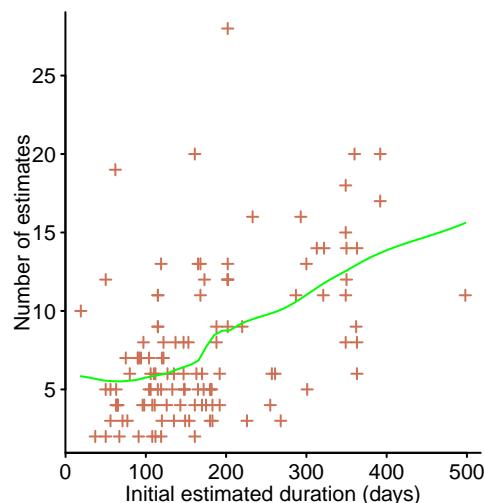


Figure 5.25: Initial estimated project duration against number of schedule estimates made before completion, for 121 projects; line is a loess fit. Data kindly provided by Little.<sup>857</sup> code

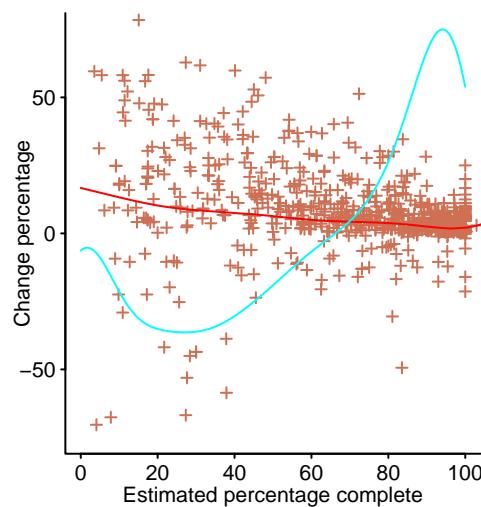


Figure 5.26: Percentage change in 882 estimated delivery dates announced at a given percentage of the estimated elapsed time of the corresponding project, for 121 projects (red is a loess fit); blue line is a density plot of percentage estimated duration when estimate made. Data kindly provided by Little.<sup>[857 code](#)</sup>

along the y-axis (red line is a loess fit). The large changes in duration tend to occur near the start of projects, with smaller changes towards the estimated end date.

The blue line is a density plot of the percentage elapsed time when a schedule change was announced (density values not shown). The flurry of new estimates making large changes the delivery date suggest that a lot of new information is discovered during project startup. Over 50% of all estimates are made in the last 71% of estimated remaining time and 25% in the last 6.4% of remaining time.

Parkinson's law claims that work expands to fill the time available. When management gives an estimated duration for a project, it is possible that those involved choose to work in a way that meets the estimate (assuming it would have been possible to complete the project in less time).

A study by van Oorschot, Bertrand and Rutte<sup>[1392](#)</sup> investigated the completion time of 424 work packages involving advanced micro-lithography systems, each having an estimated lead time. Figure 5.27 shows the number of work packages completed within a given amount of time, with colored lines showing stratification by estimated lead time.

emailed... eventually perhaps...?

Using an iterative process... time-boxed, fixed time periods, between each release...

Development groups following an Agile methodology make frequent releases, with each containing a small incremental update (Agile is an umbrella term applied to a variety of iterative and incremental software development methodologies).

Shirking, free-riding... Data promised...?

### 5.4.3 Major activities

The major activities involved in producing a system include requirements, design, implementation (or coding), testing and deployment. These activities have a natural ordering in the sense that it is unwise to deploy without some degree of testing, which requires code to test, which ideally has been designed and implemented a customer requirement.

These project activities are often called project phases, implying both an ordering and a distinct period during which one activity is performed.

A study by Zelkowitz<sup>[1483](#)</sup> investigated when work from a particular activity was performed, relative to other activities (for thirteen projects, average total effort 13,522 hours). Figure 5.28 shows considerable overlap between all activity, e.g., 31.3% of the time spent on design occurred during the coding and testing activity.

also see [projects/zelkowitz-effect.R...](#)

The extent to which documentation is a major activity, or a minor after-thought, depends on the client...

How are the available resources distributed across the major project activities?

A study by Condon, Regardie, Stark and Waligora<sup>[282](#)</sup> investigated 39 applications developed by the NASA Flight Dynamics Division between 1976 and 1992; Figure 5.29 shows ternary plots of the percentage effort, in time (red), and percentage schedule, in elapsed time (blue), for design, coding and testing (mean percentages for the three activities were: effort time: 24%, 43 and 33 respectively and schedule elapsed time: 33%, 34 and 33%).

Figure 5.30 shows a ternary plot for design/coding/test effort for projects developed by a computer manufacturer (red), telecoms company (green), space projects (blue) and major defense systems (purple).

There is a clustering of effort breakdowns for different application areas; the mean percentage design, coding and testing effort were: computer/telecoms (17%, 57, 26) and Space/Defence (36%, 20, 43). There is less scope to recover from the failures of software systems operating in some space/defense environments; this usage characteristic makes greater investment in design and testing worthwhile.

Figure 5.27: Number of work packages completed within a given time; colored lines are work packages having the same estimated lead time. Data extracted from van Oorschot et al.<sup>[1392 code](#)</sup>

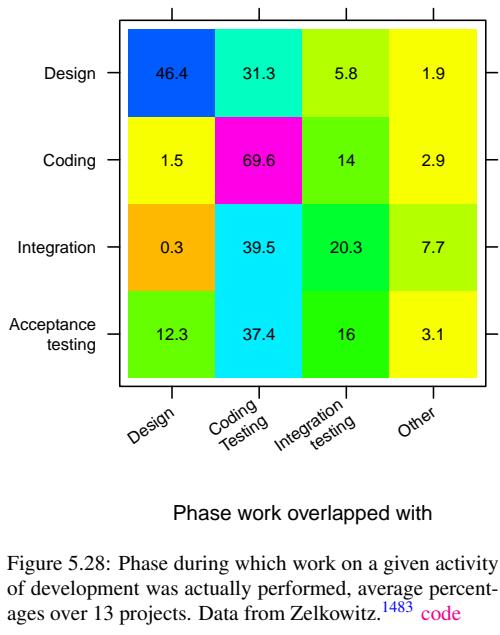


Figure 5.28: Phase during which work on a given activity of development was actually performed, average percentages over 13 projects. Data from Zelkowitz.<sup>[1483 code](#)</sup>

#### 5.4.4 Discovering functionality needed for acceptance

What functionality does a software system need to support for a project delivery to be accepted<sup>vii</sup> by the client?

The *requirements gathering* process (other terms used to describe this process include: *requirements elicitation*, *requirements capture* and *systems analysis*) will be influenced by the environment in which the project is being developed:

- when an existing way of working, or system, is being replaced (over half the projects in one recent survey<sup>712</sup>), the stakeholders of the existing system are an important source of requirements information,
- when the software is to be sold to multiple customers, as a product, those in charge of the project select the requirements. In this entrepreneurial role, the trade-off is minimising investment in implementing functionality against maximising sales income,
- when there is huge uncertainty, perhaps the client does not have enough information to decide between options or there are questions about technically feasibility, building a prototype can be a cost-effective method for evaluating ideas and technologies, helping decide and refine requirements.<sup>1275</sup>
- when starting an open source project the clients are those willing to do the work, or contribute funding.

The probability that a requirement will be added or changed, some aspect of the design has to be changed (because it is discovered that the proposed solution is unlikely to be viable), some code has to be rewritten or the hardware has to be modified...

A study by Felici<sup>422</sup> analysed the evolution of requirements, over 22 releases, for eight features contained in the software of a safety-critical avionics system. Figure 5.31 shows the requirements for some features completely changing between releases, while the requirements for other features were unchanged over many releases.

The relative cost, for different phases of development, of adding or changing a requirement, changing the design, rewriting code or modifying the hardware?...

How much effort might be required to produce a detailed requirements specification? One effort estimate<sup>699</sup> was 50 person years to produce, the C90 standard, a 219-page A4 document; the effort for the next version of the standard, C99 (a 538-page document), was estimated to be 12 person years.

What is a cost effective way of discovering requirements and their relative priority?

One possible source of requirements is the intended users of the software.

A *stakeholder* is someone who gains or loses something (e.g., status, money, change of job description) as a result of a project going live. Stakeholders are a source of political support, resistance to change<sup>1415</sup> and active subversion<sup>1178</sup> (i.e., doing what they can to obstruct progress on the project), may be able to provide essential requirements' information, or may an explicit target for exclusion (e.g., criminals with an interest in security card access systems).

Requirements gathering is an iterative process<sup>241</sup>...

Failure to identify the important stakeholders can result in missing or poorly prioritized requirements, which can have a significant impact on project success. What impact do different stakeholder selection strategies have?

A study by Lim<sup>850</sup> investigated the University College London (UCL) project to combine different access control mechanisms into one, e.g., access to the library and fitness centre. The Replacement Access, Library and ID Card (RALIC) project had been deployed two years before the study started, and had more than 60 stakeholder groups. The project documentation, along with interviews of those involved (gathering data after project completion means some degree of hindsight bias will be present), was used to create the *Ground truth* of project stakeholder identification (85 people), their rank within a role, requirements and relative priorities.

The following two algorithms were used to create a final list of stakeholders:

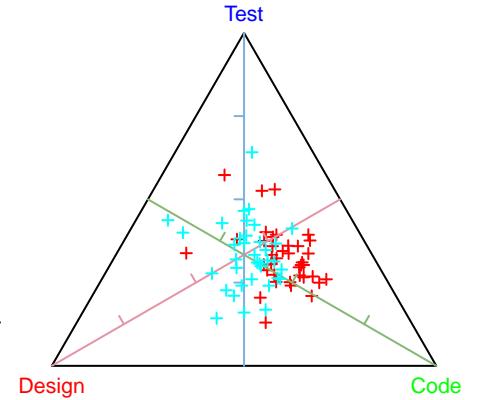


Figure 5.29: Percentage distribution of effort time (red) and schedule time (blue) across design/coding/testing for 38 NASA projects. Data from Condon et al.<sup>282</sup> [code](#)

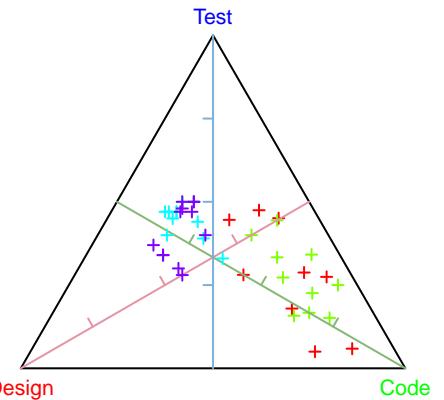


Figure 5.30: Percentage distribution of effort across design/coding/testing for 10 ICL projects (red), 11 BT projects (green), 11 space projects (blue) and 12 defense projects (purple). Data from Kitchenham et al.<sup>765</sup> and Graver et al.<sup>540</sup> [code](#)

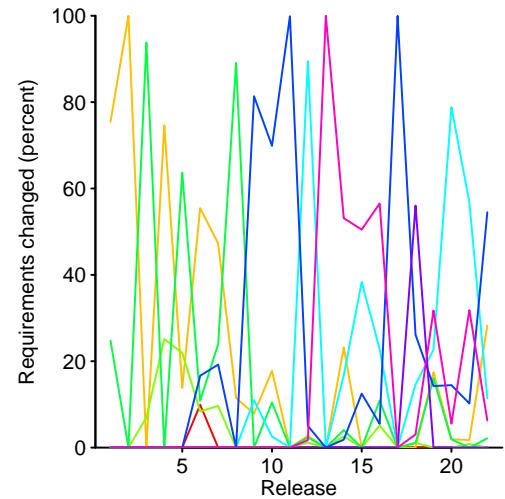


Figure 5.31: Percentage of requirements added/deleted/modified in eight features (colored lines) of a product over 22 releases. Data extracted from Felici.<sup>422</sup> [code](#)

<sup>vii</sup> Acceptance here means paying all the money specified in the contract, plus any subsequent agreed payments.

- starting from an initial list of 22 names and 28 stakeholder roles, four iterations of Snowball sampling resulted in a total of 61 responses containing 127 stakeholder names+priorities and 70 stakeholder roles (known as the *Open list*),
- a list of 50 possible stakeholders was created from the project documentation. The people on this list were asked to indicate which of those names on the list they considered to be stakeholders and to assign them a salience<sup>viii</sup> between 1 and 10, they were also given the option to suggest other names as possible stakeholders. This process generated a list containing 76 stakeholders names+priorities and 39 stakeholder roles (known as the *Closed list*).

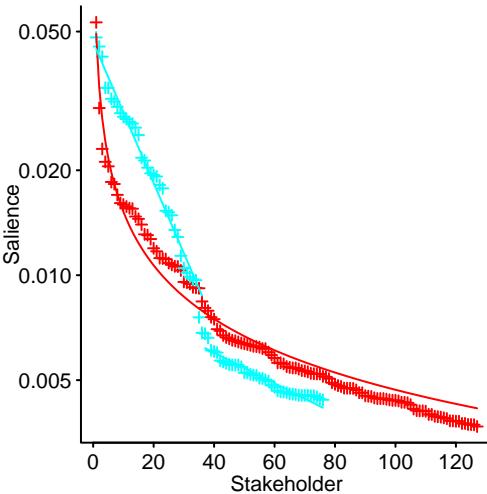


Figure 5.32: Pagerank of the stakeholders in the network created from the Open (red) and Closed (blue) stakeholder responses (values for each have been sorted). Data from Lim.<sup>850</sup> code

How might a list of people and the salience each of them assigns to others be combined to a single salience value for each person?

Existing stakeholders are in a relationship network. Lim proposed the hypothesis that the rank of stakeholder roles calculated using social network metrics would be strongly correlated with the rank ordering of stakeholder roles in the Grounded truth list. Figure 5.32 shows the Open and Closed stakeholder salience values calculated using Pagerank (treating each stakeholder as a node in the network created using the respective lists; Pagerank was chosen for this example because it had one of the strongest correlations with the Ground truth ranking). Also, see [projects/requirements/stake-ground-cor.R](#).

Identifying stakeholders and future users is just the beginning. Once found they have to be convinced to commit time to a project they may have no immediate interest in; stakeholders will have their own motivations for specifying requirements. When the desired information is contained in the heads of a few key players, these need to be kept interested and involved throughout the project. Few people are practiced in requirements' specification and so obtaining the desired information is likely to be an iterative process, e.g., they describe solutions rather than requirements.

**Prioritization** The priority of some requirements will be ranked higher than others; concentrating resources on the high priority requirements is a cost-effective way of keeping the client happy and hopefully producing a more effective system (for the resources invested). Techniques for prioritising requirements include:

- an aggregated priority list: this involves averaging stakeholders' list of assigned values, possibly weighting each stakeholders' views.

To what extent are the final requirements' priorities dependent on one stakeholder? Calculating an average, with each stakeholder excluded in turn, is one method of estimating the variance of priority assignments.

A study by Regnell, Höst, Dag, Beremark and Hjel<sup>1150</sup> asked each stakeholder/subject to assign a value to every item on the list of requirements, without exceeding a specified maximum amount (i.e., to act as-if they had a fixed amount of money to spend on listed requirements). Figure 5.33 shows the average value assigned to each requirement and the standard deviation in this value when stakeholders were excluded, one at a time.

- performing a cost/benefit analysis on each requirement and prioritizing based on the benefits provided for a given cost<sup>736</sup> (see [projects/requirements/cost-value.R](#)),
- the *analytic hierarchy process* (AHP)... data...

Figure 5.33: Average value assigned to requirements (red) and one standard deviation bounds (blue) based on omitting one stakeholder's priority value list. Data from Regnell et al.<sup>1150</sup> code

The client who authorises the cheques has the final say on which requirements have to be implemented and their relative priority. The clients' primary project role is ensuring that their desired requirements are met; clients who fail to manage the requirements process end up spending their money on a bespoke system meeting somebody else's needs.<sup>1112</sup>

## 5.4.5 Implementation

Code is an exact specification of behavior. The behaviors that matter are those that have a client visible effect, which means developers have to make many invisible decisions, e.g.,

<sup>viii</sup> The term *salience* is used to denote the level of a stakeholder's influence.

which sort routine to use. Section 6.4.3 discusses patterns of source usage. Mistakes are made during implementation and Chapter 6 discusses testing and reliability estimation.

How might project progress be measured? Lines of code is a poor metric because it is easily manipulated... i.e., Goodhart's law applies.

Percentage of time spent on activities during development... meetings, learning, coding, blocked... data...

What are the characteristics of the ongoing production of a software system developed using an iterative process?

7digital<sup>1162</sup> is a digital media delivery company that uses an Agile process to develop an open API platform providing business to business digital media services; 3,238 features were implemented by the development team between April 2009 and July 2012.

The start/done dates represent elapsed time, not total development time (i.e., work on a feature may have been stalled for a time). The measurements were not used for control or evaluation of developers, and there was no incentive for developers to change their behavior based on being measured, i.e., Goodhart's law does not apply.

Figure 5.34 shows the average time taken to implement a feature, over time (smoothed using a 25-day rolling average).

The variance in average implementation time has a change-point around 650 days (a change-point in the mean occurs around 780 days). This change suggests that one or more significant changes occurred as the second anniversary approached.

Figure 5.35 shows the number of features requiring a given number of elapsed working days for their implementation (upper first 650-days, lower post 650-days); a zero-truncated negative binomial distribution is a good fit to both sets of data.

Having a good fit to the same distribution for both the pre and post 650-day datasets suggests that the change in behavior was not a fundamental change, but akin to turning a dial on the distribution parameters one way or the other (the first 650-days has a greater mean and variance than post 650-days). If the two sets of data were better fitted by different distributions, the processes generating the two patterns of behavior are likely to have been very different.

A Negative binomial distribution suggests that the feature implementation processes has many factors involved (i.e., many values, each drawn from a different Poisson distribution, with the variation having a Gamma distribution).

Figure 5.36 shows the number of new features and number of bug fixes started per day (smoothed using a 25-day rolling mean).

7digital is a growing company, during the recording period the number of developers grew from 14 to 35, the size of its code base and number of customers is not available. The number of developers who worked on each feature was not recorded and while developers write the software, it is clients who often report most of the bugs; client information is not present in the dataset.

Possible causes for the continuing increase in new feature starts include an increase in the number of developers and/or existing developers becoming more skilled in breaking work down into smaller features (i.e., feature implementation time stays about the same because fewer developers are working on each feature, making developers available to start on new features).

Both the number of bugs and non-bug features has trended upwards, and the ratio is well below one, i.e., they spend more effort adding features than fixing bugs. Some of increase has been generated by the significant increase in number of developers over the time period, and it is also possible the group has become better at dividing work into smaller feature work items or that having implemented the basic core of the products less effort is now needed to create new features.

Work on feature implementation is very unlikely to always finish at the end of a day. Can the fit be improved by adjusting for measurement quantization?

Fitting 1,000 randomly modified half-day measurements and averaging over all results shows that the fit is slightly worse than the original data (as measured by various goodness of fit criteria): see [projects/agile-work/feat-half-day-7dig.R](#).

System dynamics model for Agile development...?

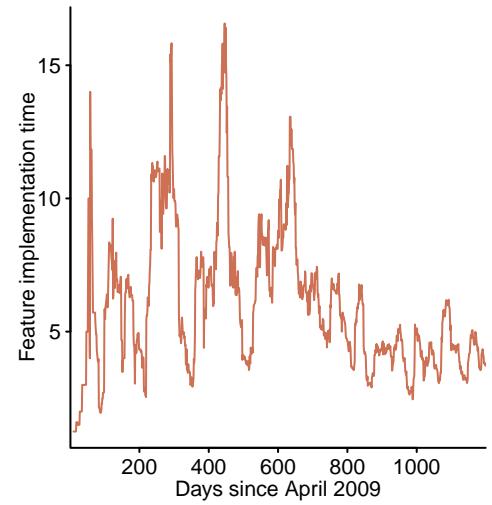


Figure 5.34: Average number of days taken to implement a feature, over time; smoothed using a 25-day rolling mean. Data kindly supplied by 7Digital.<sup>1162</sup> [code](#)

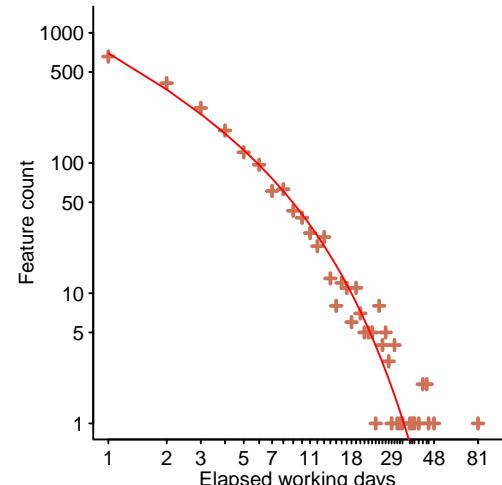
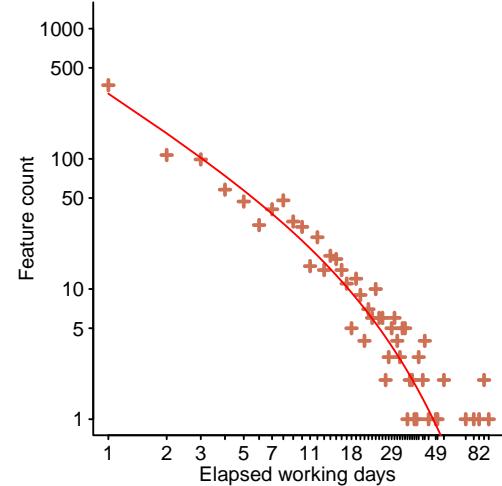


Figure 5.35: Number of features whose implementation took a given number of elapsed workdays; upper first 650-days, lower post 650-days. Fitted zero-truncated negative binomial distribution in green. Data kindly supplied by 7Digital.<sup>1163</sup> [code](#)

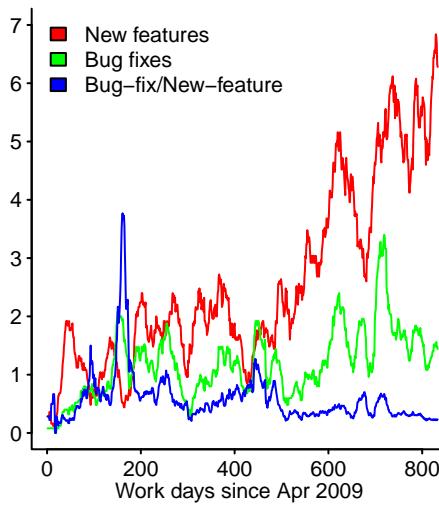


Figure 5.36: Number of feature developments started on a given work day (red new features, green bugs fixes, blue ratio of two values; 25-day rolling mean). Data kindly supplied by 7Digital.<sup>1162</sup> [code](#)

Number of merges against commits/team size/branches... broken merges...?

Work-talk patterns...??

??

## 5.4.6 Deployment

New software systems often go through a series of staged releases (e.g., alpha, then beta releases and release candidates); the intent is to uncover any unexpected problems... data on average time in various stages or release...

A successful project is likely to make multiple releases, with the number and time-frame of releases driven by the cost of making one and the likely income from users...

A study by Zhao, Serebrenik, Zhou, Filkov and Vasilescu<sup>1492</sup> investigated the impact of switching to use Travis CI as part...

Iterative development has the potential to make continuous releases, but in practice fixed (e.g., weekly) or irregular intervals...?

?

?

?

For systems supporting many build-time configurations options, it may be more cost effective<sup>572</sup> to concentrate on the popular option combinations and wait until problems with other configurations are reported (rather than invest resources checking many options that will never be used).

While some bespoke software is targeted at particular computing hardware, long-lasting software may be ported to a variety of different platforms. The cost/benefit analysis of investing in reducing the cost of porting to a different platform (i.e., reducing the switching cost) requires knowing the likelihood of this event occurring.

A study by Peukert<sup>1082</sup> investigated the switching costs of outsourced IT systems, as experienced by U.S. Credit Unions. Figure 5.37 shows the survival curve of IT outsourcing suppliers employed by 2,382 Credit Unions, over the period 2000 to 2010.

Given sufficient customer demand, it is worthwhile for a vendor to continue maintaining a product. Adding new features to an existing product creates something that can be sold to existing customers and helps to compete with new market entrants.

Figure 5.38 shows the growth of lines of code, command line options, words in the manual and messages supported by PC-Lint (a C/C++ static analysis tool), in the 11 major releases over 28 years.

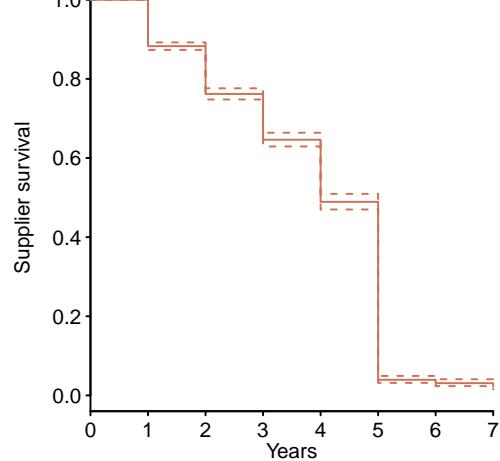


Figure 5.37: Survival curve of IT outsourcing suppliers continuing to work for 2,382 Credit Unions. Data kindly provided by Peukert.<sup>1082</sup> [code](#)

## 5.4.7 Maintenance

The maintenance of a software system is a, potentially long term, project in its own right.

Why maintain software?

Chapter 4 discusses one reason for maintaining commercial products, the output from maintenance activities is a potential source of revenue after the initial sale and a signal to potential customers that the product is not dead; see [ecosystems/maint-dev-ratio.R](#).

A company developing bespoke software for in-house may want the software to be adapted, as the world in which it is used changes. Developers are needed to fix faults and provide support. What else is there to do with their time? Adding features is a hedonic incentive for developers to stay in their current job (the issue of attracting developers to perform maintenance work on existing software is discussed)... data?...

Once funding for an academic research project has ended, maintenance of software written for that project may cease. A study<sup>646</sup> of 214 packages associated papers published between 2001-2015, in the journal Molecular Ecology Resources, found that 73% had not been updated since publication.

?

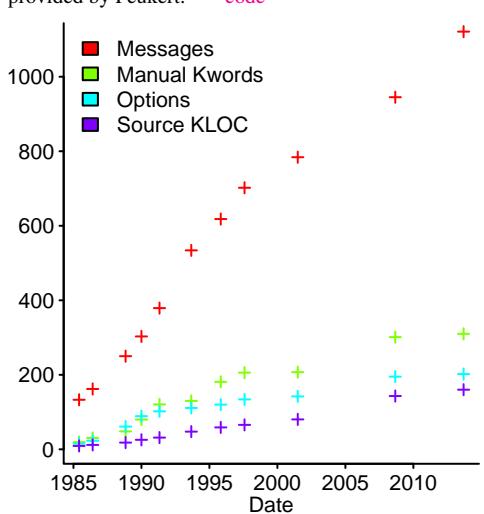


Figure 5.38: Growth, over 11 major releases in 28 years, of messages supported, command line options, kilo-words in product manual and thousand lines of code in PC-Lint. Data kindly provided by Gimpel.<sup>499</sup> [code](#)

To what extent might library dependencies be updated for new releases of a software system?...?

Some very large systems grow through the integration independently developed projects.

A study by Bavota, Canfora, Di Penta, Oliveto and Panichella<sup>109</sup> investigated the growth of projects in the Apache ecosystem. Figure 5.39 shows the growth in the number of projects, and coding constructs they contain. The number of dependencies between projects increases as the number of projects increases, along with the number of methods and/or classes (see [ecosystems/icsm2013\\_apache.R](#)).

A study by Bavota, Canfora, Di Penta, Oliveto and Panichella<sup>110</sup> investigated 147 projects in the Apache ecosystem. Figure 5.40 shows, for each pair of projects, the percentage overlap of developers contributing to both projects (during 2013; white none, red 100%); the rows/columns have been reordered to show project pair clusters sharing a higher percentage of developers (see [«\\_seriation>](#)).

## 5.5 Development teams

A project needs people to do the work, and the complexity of staff scheduling may range from a selecting a fixed number of staff to work on a project until delivery, to balancing staffing needs across multiple large projects in multiple locations.

A study by Buettner<sup>198</sup> investigated large software intensive systems and included an analysis of various staffing related issues, including: staffing level over time (Figure 11.65), staff work patterns (Figure 11.57) and impact of staffing on fault detection (Figure ??).

Distribution of number of people in a team... data?...

When do team members come from?... existing in-house staff, hire developers for the duration of the project (e.g., contractors)...

Changing staffing requirements and changing skill required staff as project progresses...

What set of skills does a development team require and how might these skills be divided up among team members?...

What is the best way to organize a software project team?

Drawing a parallel with the methods of production used in manufacturing factories, the factory concept for software projects<sup>305</sup> was proposed and used by several large companies. <sup>ix</sup> The perceived advantages of this approach are the same as those it provides to traditional manufacturers, e.g., control of the production process and reduction in the need for highly skilled employees.

There have been few experimental comparisons<sup>1456</sup> of the many project development techniques that have been proposed over the years.

The chief programmer team<sup>88</sup> approach to team organization was designed to handle environments where many of the available programmers are inexperienced (a common situation in a rapidly growing field); an experienced developer is appointed as the chief programmer and is responsible for doing the heavy lifting and allocating the tasks requiring less skill to others. This form of team organization dates from the late 1960s, when programming involved a lot of clerical activity and in its original formulation emphasis is placed on delegating this clerical activity,

Team members' awareness of other members of the team?...

Self-organising teams is one approach to dividing up the available manpower...

For some applications there is a commercial incentive for companies to cooperate to build a common system, e.g., sharing in a winner take-all market or because the benefits of owning an in-house system are not worth the costs. The projects to build these applications have teams containing developers from multiple companies...

A study by Teixeira, Robles and González-Barahona<sup>1342</sup> investigated the evolution of developers, employed by a given company, working in the OpenStack project between 2010 and

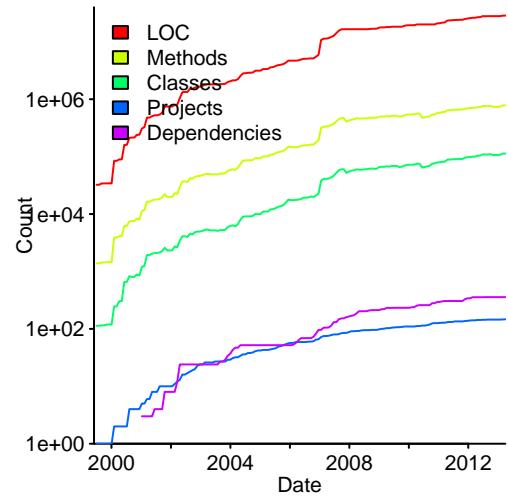


Figure 5.39: Growth in the number of projects within the Apache ecosystem, along with the amount of contained code. Data from Bavota et al.<sup>109</sup> [code](#)

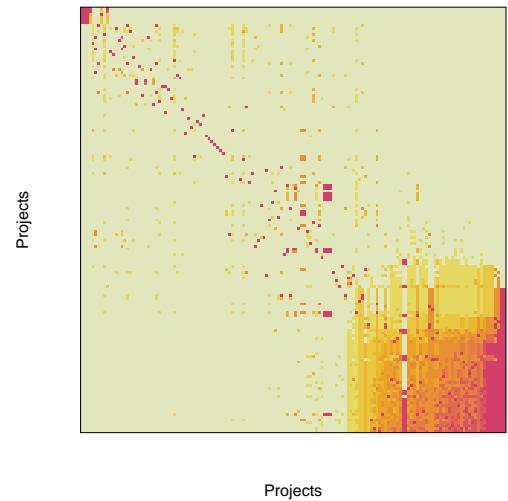


Figure 5.40: Percentage overlap of developers contributing, during 2013, to both of each pair of 147 Apache projects. Data kindly provided by Panichella.<sup>110</sup> [code](#)

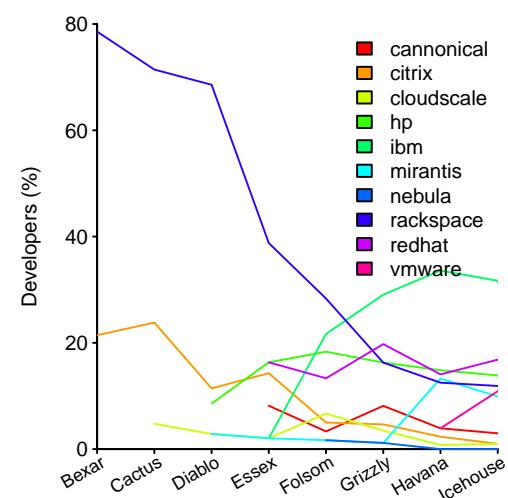


Figure 5.41: Percentage of developers, employed by given companies, working on OpenStack at the time of a release (x-axis). Data from Teixeira et al.<sup>1342</sup> [code](#)

<sup>ix</sup> It was particularly popular in Japan. Your author has not been able to locate any data on companies recently using the factory concept to produce software.

2014. Figure 5.41 shows the involvement of top ten companies, out of over 200 organizations involved, as a percentage of employed developers working on OpenStack.

Application domain knowledge<sup>1229</sup>...?

??

### 5.5.1 New staff

New people may join a project as part of planned growth, the need to handle new work, existing people leaving, management wanting to reduce the dependency on a few critical people, and many other reasons.

Project member turnover... joining/leaving data...

Training people (e.g., developers, documentation writers) who are new to a project reduces the amount of effort available to building the system in the short term. Training is an investment in people whose benefit is the post-training productivity these people bring to a project.

Brooks' Law<sup>190</sup> says: "Adding manpower to a late software project makes it later", but does not say anything about the impact of not adding manpower to a late project. Under what conditions does adding a person to a project cause it to be delayed?

If we assume a new person diverts, from the project they join, a total effort,  $T_e$ , in training and that after  $D_t$  units of time the trained person contributes  $E_n$  effort per unit time until the project deadline; unless the following inequality holds, training a new person results in the project being delayed:

$$E_{a1}D_r < (E_{a1}D_t - T_e) + (E_{a2} + E_n)(D_r - D_t)$$

where:  $E_{a1}$  is the total daily effort produced by the team before the addition of a new person,  $E_{a2}$  the total daily effort produced by the original team after the addition and  $D_r$  is the number of units of time between the start of training and the delivery date/time.

Adding a person to a team can both reduce the productivity of the original team (e.g., by increasing the inter-person communication overhead) and increase their productivity (e.g., by providing a skill that means the whole is greater than the sum of its parts). Assuming that  $E_{a2} = cE_{a1}$ , the equation simplifies to:

$$T_e < (D_r - D_t)(E_n - (1 - c)E_{a1})$$

and...

In practice a new person's effort contribution ramps up from zero, perhaps even during the training period, to a relatively constant long term daily average;  $E_n$ ...

The effort,  $E_T$ , that has to be invested in training a new project member will depend on their existing level of expertise with the application domain, tools being used, coding skills, etc (pretty much everything was new, back in the day, for the project analysed by Brooks, so  $E_T$  was probably very high). There is also the important ability, or lack of, to pick things up quickly, i.e., their learning rate...

If many people are being added to a project at the same time, it is easy to imagine it grinding to a halt because of all the minor congestion that occurs within the network of dependencies that project progress is waiting on.

In practice new developers have to learn directly from the source code.... Risks of out-of-date documentation...

# Chapter 6

## Reliability

### 6.1 Introduction

People are willing to continue using software which has faults, that they sometimes experience,<sup>i</sup> provided it delivers a worthwhile benefit to them. The random walk of life can often be nudged to avoid unpleasantness, or the operational usage time limited to keep within acceptable safety limits.<sup>185</sup> Regions of acceptability may exist in programs containing many apparently major mistakes, supporting useful functionality.<sup>1161</sup>

Software systems containing likely fault experiences are shipped because it is not economically worthwhile fixing some of the mistakes made during their implementation; also, the process of finding and fixing problems, prior to release, is often constrained by available manpower and marketing deadlines.

Software release decisions involve weighing whether the supported functionality provides enough benefit to be attractive to customers (i.e., they will spend money to use it), after factoring in likely costs arising from faults experienced by customers (i.e., from lost sales, dealing with customer complaints and possible fixing problems and making available an updated version).

How many fault experiences will customers tolerate, before they are unwilling to use software; are some kinds of fault experiences more likely to be tolerated than others? Customer utility function... Willingness to pay is a commonly used measure of risk acceptability and *As Low As Reasonably Practicable* (ALARP) is a term often applied...

Developers make the coding mistakes that create potential fault experiences, and the environment in which the code executes provides the input that results in faults occurring (which may be experienced by the user). This chapter discusses the kinds of mistakes made, where they occur in the development process, methods used to locate them and techniques for estimating how many fault experiences can potentially occur. Issues around the selection of algorithms is outside the scope of this chapter; algorithmic reliability issues include accuracy<sup>341</sup> and stability of numerical algorithms,<sup>615</sup> and solutions include minimising the error in a dot product by normalizing the values being multiplied.<sup>384</sup>

People make mistakes,<sup>1123,1148</sup> economic considerations dictate how much is invested in reducing the probability that mistakes costly leading to costly fault experiences remain (either contained in delivered software systems or as a component of a larger system). The fact that programs often contained many mistakes was a surprise to the early computer developers,<sup>1453</sup> as it is for people new to programming.

Operating as an engineering discipline, does not in itself ensure that a constructed system has the desired level of reliability. There has been roughly a 30-year cycle for bridge failures;<sup>1081</sup> new design techniques and materials are introduced, and growing confidence in their use leads to overstepping the limits of what can safely be constructed.

What constitutes reliability, in a given context, is driven by customer requirements, e.g., in some situations it may be more desirable to produce an inaccurate answer than no answer at all, while in other situations no answer is more desirable than an inaccurate one.

---

<sup>i</sup> Experience is the operative word, a fault may occur and not be recognized as such.

The early computers were very expensive to buy and operate, and much of the software written in the 1960s and 1970s was for large corporations or government bodies; the US Department of Defence took an active role in researching software reliability and much of the early published research is based on the software development environments used by DOD and NASA projects during this period.

The approach to computer system reliability promoted by these sponsors of early research set the outlook for much of what has followed. Emphasis is often placed either on large projects that are expected to be maintained over many years or software operating in situations where the cost of failure is extremely high and there is very limited time, if any, to fix problems (e.g., Space shuttle missions<sup>516</sup>).

This chapter discusses reliability from a cost/benefit perspective; yes, mistakes have a cost, but these can be outweighed by the benefits of releasing the software containing them. As with the other chapters, the target audience for this material is the vendor of the software, not its users; it is possible for vendors to consider a project a success because it produced a good return on investment, but for the user to consider it to be a failure because of the problems they experienced using it.

The relative low cost of modifying existing software, compared to hardware, provides greater flexibility for trading-off upfront costs against the cost of making changes later (e.g., by reducing the amount of testing before release), knowing that if necessary it is often practical to provide updates later. For some vendors, the Internet provides an almost zero cost means of distribution.

Mistakes in software can have a practical benefit for some people, for instance authors of computer malware have used mistakes in CPU emulators to detect that their activity may be monitored<sup>1047</sup> (and therefore the malware should remain inactive).

Mistakes are not unique to software systems; a study of citations in research papers<sup>937</sup> found an average error rate of 20%.

The creation of mathematics shares many similarities with writing software and plenty of mistakes are made in mathematics.<sup>1106</sup> The size, complexity and technicality of some mathematical proofs has now reached the point where serious questions are being raised about the ability of anybody to check whether they are correct, e.g., Mochizuki's proof of the *abc* conjecture<sup>221</sup> and the Hales-Ferguson proof of the *Kepler Conjecture*.<sup>801</sup> Many important theorems don't have proofs, only sketches of proofs and outline arguments that are believed to be correct;<sup>994</sup> the sketches provide evidence used by other mathematicians to decide whether they believe a theorem is true (a theorem may be true, even although mistakes are made in the claimed proofs).

The social processes involved in the mathematics community coming to believe that a theorem is true, is still coming to terms with believing machine-checked proofs.<sup>1100</sup> The nature and role of proof in mathematics continues to be debated.<sup>609</sup>

Mistakes are much less likely to be found in mathematical proofs than software, because a lot of specialist knowledge is needed to check new theorems in specialised areas, but a clueless button pusher can experience a fault in software simply by running it; also, there are few people checking proofs, while software is being checked every time it is executed.

Proposals<sup>623</sup> that programming should strive to be more like mathematics is based on the misconception that the process of creating proofs in mathematics is less error prone than creating software.

Software that has been both developed and checked using formal analysis tools may still contain mistakes;<sup>453,488</sup> there is also the possibility of mistakes and ambiguities in the tools used to perform the verification, e.g., ML.<sup>723</sup> One study<sup>972</sup> of a software system that had been formally provided to be correct, found at least two mistakes per thousand lines remained.

Software differs from hardware in that, not only does it not wear out, but the variability in the inputs it has to process can have a significant impact on fault experiences. In software, repeatedly performing the same operation with the same input is expected to produce the same behavior (while hardware is expected to eventually fail). The dependency of fault experiences on the input distribution means much of the hardware-based reliability theory is not applicable.

Fixing a reported fault is one step in a chain of events that may result in users of the software receiving an update.

Vendors who supply an operating system vary in the amount of control they exercise over the software version their customers can run. Apple maintains a tight grip over use of iOS, and directly supplies updates to customers cryptographically signed for a particular device (i.e., the software can only be installed on the device that downloaded it). Google supplies the latest version of Android to OEMs and has no control over what, if any, updates these OEMs supply to customers (who are also free to install versions from third-party suppliers). Microsoft sells Windows 10 through OEMs, but makes available security fixes and updates for direct download by customers.

Figure 6.1 shows some of the connections between participants in the Android ecosystem (number of each kind in brackets), and some edges are labeled with the number of known updates flowing between particular participants (from July 2011 to March 2016).

Experiments designed to uncover unreliability issues may fail to find any. This does not mean that they are rare, the reason for failing to find a problem may be lack of statistical power (i.e., the likelihood of finding an effect if one exists); this topic is discussed in Section 10.6.

### 6.1.1 It's not a fault, it's a feature

The classification of program behavior as a fault or a feature may depend on the person doing the classification (e.g., user or developer). For instance, software written to manage a parts inventory may not be able to add a new part once the number of parts it contains equals 65,536; a feature/fault that users will not encounter until the number of parts in the inventory reaches this value.

One fault report study<sup>610</sup> found that many reported faults were actually requests for enhancement.

Is insufficient or excessive accuracy in the numeric values calculated by a program, a fault or feature?

Variations in the behavior of software between different releases, or running the same code on different hardware can be as large as the behavior effect the user is looking for; a potential major problem when medical diagnosis is involved<sup>556</sup> ...

The accuracy of calculated results may be specified in the requirements or the developer writing the code may be the only person who gives any thought to the question. When calculations involve floating-point values, developers may choose to play safe and use variables declared to have types capable of representing greater accuracy than is required<sup>1186</sup> (leading to higher than necessary resource usage,<sup>537</sup> e.g., memory, time and battery power).

Effects of using fewer representation bits in a floating-point datatype<sup>1338</sup>

Hardware design choices can increase the difficulty of obtaining accurate numeric results, e.g., the choice of representation for numeric values. Cost of implementation<sup>298</sup> resulted in binary becoming the dominant hardware representation for numeric values, despite decimal being the dominant representation within business. While cost of implementation eventually decreased to a point where processors supporting both binary and decimal are generally available, much existing software is written from a binary perspective.

### 6.1.2 Why do faults occur?

Two events are required for the person running an application to experience a software related fault:

- a mistake exists in the software, i.e., if the mistake is found, the developer who wrote the code would change it to behave as intended,
- the program processes input values that cause it to execute the code containing the mistake in a way that results in a fault being experienced<sup>295</sup> (software that is never used has no reported faults).

Some mistakes are more likely to be encountered than others because the required input values are more likely to occur during the use of the software. Any analysis of software reliability has to consider the interplay between the probabilistic nature of the input distribution and coding mistakes present in the source code (or configuration information).

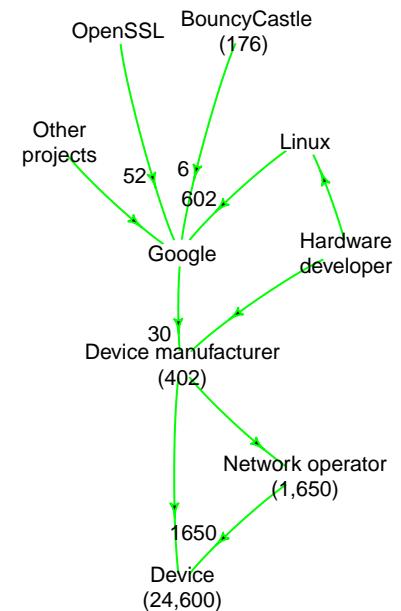


Figure 6.1: Flow of updates between participants in one Android ecosystem; number of each kind of member given in brackets, number of updates shipped on edges. Data from Thomas.<sup>1351</sup> code

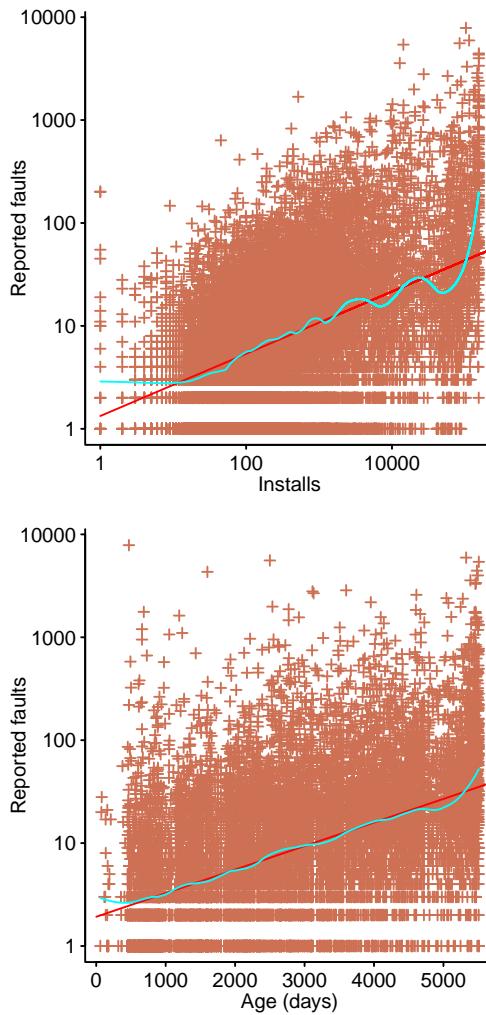


Figure 6.2: Reported faults against number of installations (upper) and age (lower). Data from the "wheezy" Debian release.<sup>1376</sup> code

An increase in the number of people using a program is likely to lead to an increase in reported faults, because of both an increase in possible reporters and an increase in the diversity of input values.

The Ultimate Debian Database project<sup>1376</sup> collects information about packages included in the Debian Linux distribution, from users who have opted-in to the Debian Popularity Contest. Figure 6.2 shows the numbers of installs (for the "wheezy" release) of a given packaged application against faults reported in that package and also age of the package against faults (data from the Debian Bug Tracking System, which is not the primary fault reporting system for some packages); also, see Figure 11.23.

A fitted regression model is, assuming a percentage measurement error:

$$\text{reported\_bugs} = e^{-0.15 + 0.17 \times \log(\text{insts}) + (30 + 2.3 \times \log(\text{insts})) \times \text{age} \times 10^{-5}}$$

For an *age* between 1,000—6,000 and installs between 10—20,000 (giving a  $\log(\text{insts})$  between 2—10), then the number of installations (a proxy for number of users) appears to play a larger role in the number of reported faults, compared to *age* (i.e., the amount of time the package has been included in the Debian distribution). There is a great deal of variance in the data...

A study<sup>1319</sup> of TCP checksum performance found that a far fewer corrupt network packets were detected in practice, than expected (by a factor of between 10 and 100). The difference between practice and expectation was caused by the non-uniform distribution of input values (the proof that checksum values are uniformly distributed assumes a uniform distribution of input values).

A hardware fault may cause the behavior of otherwise correctly behaving software to appear to be wrong, e.g., hardware is more likely to fail as the workload increases.<sup>675</sup>

### 6.1.3 Reported fault data

While reported faults have been studied almost since the start of software development, until open source bug repositories became available there was little publicly available fault data. The possibility of adverse publicity and fear of legal consequences of publishing information on problems found their software products was not lost on commercial organizations and nearly all of them treat such information as commercially confidential. While some companies maintained software fault databases,<sup>534</sup> these were not publicly available.

During the 1970s the Rome Air Defence Center published many detailed studies of software development,<sup>304</sup> and some included data on faults experienced in military projects.<sup>1458</sup> However, these reports were not widely known about or easy to obtain, until they became available via the Internet; a few were published as books.<sup>1349</sup>

The few pre-Open source datasets appearing in academic papers contained relatively few fault reports and if submitted for publication today would probably be rejected as not worthy of consideration. These studies<sup>1077, 1079</sup> usually investigated particular systems, listing percentages of faults found by development phase and the kinds of faults found; one study<sup>1421</sup> listed fault information relating to one application domain, medical device software. Knuth published<sup>775</sup> a list of all known mistakes in a widely used program (LaTex).

Confidentiality, unwillingness to share data, researchers not archiving their data once work on a project stops...

The economic impact of poor computer security has resulted in coding mistakes having security implications becoming a major category of interest. Several databases of such faults are actively maintained, including: The NVD (National Vulnerability Database<sup>1014</sup>), the VERIS Community Database (VCDB),<sup>1397</sup> an Open source vulnerability database, the Exploit database (EDB)<sup>373</sup> lists proof of concept vulnerabilities; mistakes in code which may be exploited to gain unauthorised access to a computer (vulnerabilities discovered by security researchers who have a motivation to show off their skills), and the Wooyun program<sup>1490</sup>...

A study by Sadat, Bener and Miranskyy<sup>1191</sup> investigated issues involving connecting duplicate fault reports. Figure 6.3 shows the connection graph for Eclipse report 6325 (report 4671 was the earliest report covering this issue).

How similar are the characteristics of Open source project fault report data compared with commercial fault data?

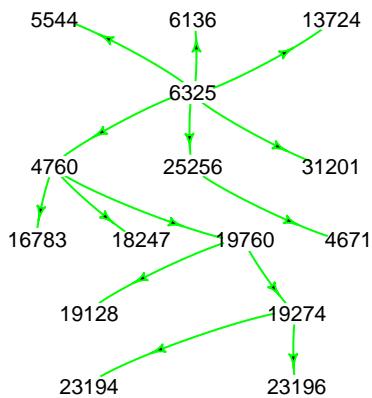


Figure 6.3: Duplicates of Eclipse fault report 4671 (report 6325 was finally chosen as the master report); arrows point to report marked as duplicate of an earlier report. Data from Sadat et al.<sup>1191</sup>

Various problems have been found with Open source fault report data, which is not to say that fault data on closed source projects is not without its own problems; these problems include:

- reported faults do not always appear in the fault report databases (e.g., serious bugs tend to be under-reported in commit logs;<sup>142</sup> a replication<sup>1006</sup>). One study<sup>74</sup> of fault reports for Apache, over a 6-week period, found that only 48% of bug fixes were recorded as faults in the Bugzilla database; the working practice of the core developers was to discuss serious problems on the mailing list and many fault experiences were never formally logged with Bugzilla,
- reported faults are misclassified. One study of reported faults (see [?]) found that 42.6% of fault reports had been misclassified, with 39% of files marked as defective not actually containing any reported fault... requests for enhancement... Differences of opinion in how a reported fault should be classified<sup>1402</sup>... reported problem, on detailed investigation, turning out not to be a coding mistake,
- reporting bias: reported faults discovered through actively searching for them,<sup>410</sup> rather than normal program usage (e.g., Linux is a popular target for researchers, fuzzing tools; csmith generated source code targeted at compilers). The reporting of vulnerabilities contained, or not contained, in NVD has been found to be driven by a wide variety social, technical and economic pressures.<sup>254, 1007</sup> Data on faults discovered through an active search process may have different characteristics than faults experienced through normal program usage,
- reported faults are not contained within the source of the program cited, but is contained within a third-party library: a study by Ma, Chen, Zhang, Zhou and Xu<sup>876</sup> of the Python scientific package ecosystem found...
- reported fault could not be reproduced or was intermittent: a study<sup>1193</sup> of six servers found that on average 81% of the 266 reported faults analysed could be reproduced deterministically, 8% non-deterministically, 9% were timing dependent, plus various other cases,

A lot of fault analysis research is based on the faults reported against brand-name systems, such as Linux, Mozilla, Eclipse and NVD...

The fault report data does not always contain enough information to answer the questions being asked of it, e.g., using incidence data to distinguish between different exponential order fault growth models.<sup>949</sup>

#### 6.1.4 Cultural outlook

Cultures vary in their attitude to the risk of personal injury and death. A study by Viscusi and Aldy<sup>1412</sup> investigated the value of a statistical life in 11 countries and found a range of estimates from \$0.7 million to \$20 million (adjusted to the dollar rate in 2000). Individuals in turn have their own perception of risk and sensitivity to the value of life.<sup>1273</sup> Some risks may be sufficiently outside of a persons' experience that they are willing to discount an occurrence affecting them, e.g., destruction of a city, country or all human life by a meteor impact.<sup>1279</sup>

Linguistic studies of concepts associated with the word *reliability* have been made for Japanese<sup>955</sup>

A study by Budescu, Por, Broomell and Smithson<sup>197</sup> investigated how people in 24 countries, speaking 17 languages, interpreted uncertainty statements containing four probability terms (i.e., very unlikely, unlikely, likely and very likely, translated to the subjects' language). Figure 6.4 shows the mean percentage likelihood estimated by people in each country to statements containing each term.

Public perception of events influences and is influenced by media coverage (e.g., in a volcano vs. drought disaster, the drought needs to kill 40,000 times as many people as the volcano to achieve the same probability of media coverage<sup>380</sup>). A study<sup>380</sup> of disasters and media coverage found that when a disaster occurs while other stories are deemed more newsworthy, aid from U.S. disaster relief is less likely to occur.

Table 6.1, from a 2011 analysis by the UK Department for Transport,<sup>1359</sup> lists the average value that would have been saved, per casualty, had an accident not occurred.

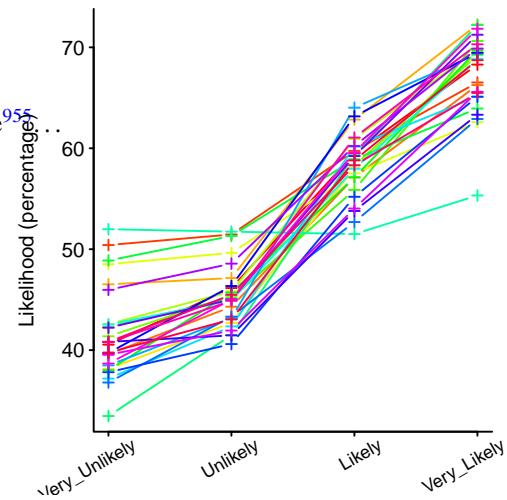


Figure 6.4: Mean percentage likelihood of (translated) statements containing a probabilistic term; one colored line per country. Data from Budescu et al.<sup>197</sup> code

Injury severity	Lost output	Human costs	Medical and ambulance	Total
Fatal	£545,040	£1,039,530	£940	£1,585,510
Serious	£21,000	£144,450	£12,720	£178,160
Slight	£2,220	£10,570	£940	£13,740
Average	£9,740	£35,740	£2,250	£47,470

Table 6.1: Average value of prevention per casualty, by severity and element of cost (human cost based on willingness-to-pay values); last line is average over all casualties. Data from UK Department for Transport.<sup>1359</sup>

A study by Costa and Kahn<sup>293</sup> investigated changes in the value of life in the USA, between 1940 and 1980. The range of estimates, adjusted to 1990 dollars, was \$713,000 to \$996,000 in 1940 and \$4.144 million to \$5.347 million in 1980.

Estimates of the number of deaths associated with computer related accidents contain a wide margin of error,<sup>883</sup> with human-computer interaction being a commonly cited cause...

Governments are aware of the dangers of society becoming overly risk-averse, and some have published risk management policies<sup>1419</sup>...

A report<sup>252</sup> from the UK's Department for Environment, Food and Rural Affairs provides an example of the kind of detailed analysis involved in calculating a monetary valuation of reducing risk.

The U.S. Department of Defense Standard MIL-STD-882E<sup>343</sup> defines and gives numbers for the terms, when applied to an individual item; the following are some of these terms:

- *Probable*: "Will occur several times in the life of an item"; probability of occurrence less than  $10^{-1}$  but greater than  $10^{-2}$ .
- *Remote*: "Unlikely, but possible to occur in the life of an item"; probability of occurrence less than  $10^{-3}$  but greater than  $10^{-6}$ .
- *Improbable*: "So unlikely, it can be assumed occurrence may not be experienced in the life of an item"; probability of occurrence less than  $10^{-6}$ .

Some government related organizations have published guidelines covering the use of software in various industries, e.g., in medical devices,<sup>415</sup> automotive<sup>957,958</sup>...

## 6.2 The search for profit

A vendor's approach to product reliability is driven by maximising return on investment in their chosen market.

There is often an asymmetry in the cost of implementation mistakes borne by the producer and consumer of software systems.

For single client development projects, the client is in a position to make trade-off decisions involving their estimated post-delivery fault experience costs and estimated pre-delivery mistake reduction costs (e.g., paying the vendor to do something).

For software written to be sold to multiple customers, the producer has to cover the costs for their mistakes (which may include costs arising from litigation<sup>733</sup>) and the customer...

The reputational cost, to the vendor, of customer fault experiences...

In some markets the likelihood of customer fault experiences is source of revenue to vendors, e.g., the profit margin on a maintenance contract....

In some markets, there may be a competitive advantage to being first to market with a new or updated product; the biggest component of fault experience costs may be lost sales rather than the cost of correcting mistakes later (i.e., take the risk that customers are not deterred by the greater number of faults experienced)... rapid development, *frontier risk* thinking...

Mistakes in code may cease to exist because the code containing them is rewritten or the hardware on which the software runs is replaced. Correcting a coding mistake that is unlikely to result in a customer fault experienced is a wasted investment.

A study by Di Penta, Cerulo and Aversano<sup>349</sup> recorded the output from running various static analysis tools (Rats, Splint and Pixy) on each release of several large software systems (i.e., Samba, Squid and Horde). Comparing the tool output from each software release, it was possible to find the first/last release where a warning occurred for a particular line of code (or may still be occurring in the latest release).

Figure 6.5 shows the survival curve for the two most common warnings reported by Splint (memory problem and type mismatch, over 85% of all generated warnings). In the case of Splint type mismatches are more likely to be removed before memory problems first while for Squid memory problems are much more likely to be removed before type mismatches; also see Figure 11.75.

The average lifetime of coding mistakes varies between programs and kind of mistake.<sup>1048</sup>

Some hardware devices have a relatively short lifetime, e.g., mobile phones and graphics cards. Comparing the survival rate of reported faults in Linux device drivers and other faults in Linux shows that for the first 18 months, or so, the expected lifetime of reported faults in device drivers is much shorter than faults in other systems (see Figure 6.6); thereafter, the two faults lifetimes are roughly the same...<sup>1049</sup>

Also, see:<sup>?</sup>

emailed for fig 2 data...<sup>1351</sup>?

Files ported from forked versions of BSD contain fewer mistakes...<sup>1144</sup> emailed...

While coding mistakes are exploited by computer viruses, causing business disruption, the greatest percentage of computer related losses come from financial fraud by insiders and traditional sources of loss such as theft of laptops and mobiles.<sup>1158</sup>

All mistakes have the potential to have costly consequences, but in practice most appear to be an annoyance. One study<sup>23</sup> found that only 2.6% of the vulnerabilities listed in the NVD have been used, or rather their use has been detected, in viruses and network threat attacks on computers.

emailed for data...<sup>?</sup>

A market has developed for faults that allow third parties to gain control of other peoples' computers<sup>22</sup> (e.g., spying agencies and spammers), and some vendors have responded by creating vulnerability reward programs. The list of published rewards are a measure of the value vendors place on this particular kind of fault experience. plot values<sup>435</sup> and<sup>?</sup>...

May be more cost effective to pay bug bounties than investing in finding mistakes prior to release...<sup>435</sup>...

Over half of those registered with a bounty program report a single bug, a quarter report two,<sup>888</sup> also see Figure 4.14...

The *willingness-to-pay* approach attempts to determine the maximum amount that those at risk would individually be willing to pay for improvements to their or other people's safety. Each individual may only be willing to pay a small amount, but if they are a member of a group the amounts can be added together to estimate a value for the group "worth" of a safety improvement. For instance, assuming that in a group of 1.5 million people, each person is willing to pay £1 for safety improvements that achieve a 1 in 1 million reduction in the probability of death; the summed WTP *Value of Preventing a Statistical Fatality* (VPF) is £1.5 million (the same approach can be used to calculate *Value of Preventing non-fatal Injuries*).

If a customer reports a fault, in software they have purchased, what incentive does the vendor have to correct the problem and provide an update (they have the customers' money; assuming the software is not so fault ridden that it is returned for a refund)? Possible reasons include:

- it would be more expensive not to fix the coding mistake, e.g., the change in behavior caused by the fault experience could cause an accident, or negative publicity, that has an economic impact greater than the cost of fixing the mistake. Not wanting to lose money because a mistake had consequences that resulted in legal action...
- a customer support agreement requires certain kinds of reported faults to be fixed...

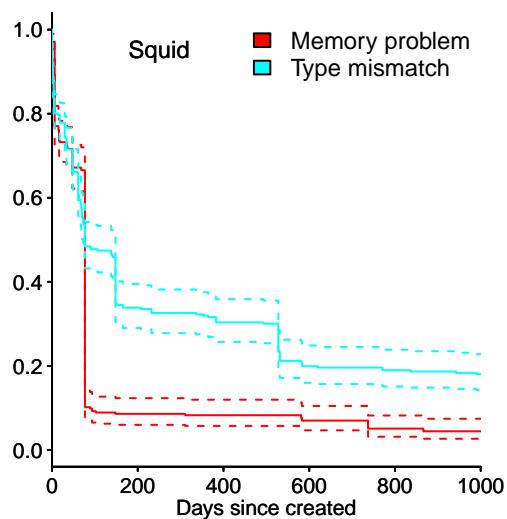
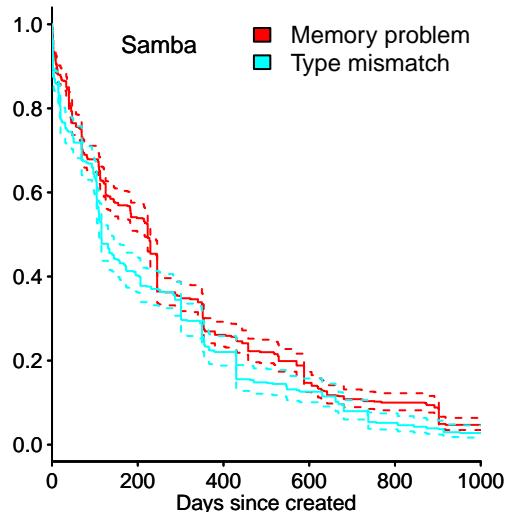


Figure 6.5: Survival curve of the two most common warnings reported by Splint in Samba and Squid. Data from De Penta et al.<sup>349</sup> [code](#)

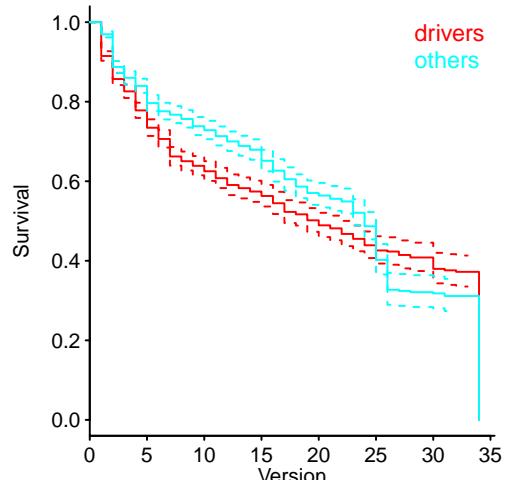


Figure 6.6: Survival rate of reported faults in Linux device drivers and other Linux subsystems... Data from Palix et al.<sup>1049</sup> [code](#)

- public perception and wanting to maintain customer good will, in the hope of making further sales.

The time taken to fix publicly disclosed vulnerabilities is shorter than for vulnerabilities disclosed privately to the vendor; see the analysis in Section 11.11.3.1,

- it is fill-in work for developers between product releases...

Which implementation mistakes are corrected? While there is no benefit in correcting mistakes that customers are unlikely to experience, it may not be possible to deduce whether it is likely to produce a customer fault experience and so it has to be added to the list of known problems. Once a product has been released and known to be acceptable to many customers there is not be any incentive to actively search for potential fault experiences, consequently the only mistakes likely to be corrected are the ones reported by customers.

The vendor decision on whether to handle a reported fault is based on a cost/benefit trade-off involving the following factors:

- does a simple workaround exist, so users can continue using the software without needing an update, i.e., is it possible to reduce the issue to a low priority,
- the likely cost of correcting the mistake,
- the probability that a customer will experience a fault and recognize it as such,
- the impact to the vendor if a known mistake is not corrected. Adverse publicity when cost/benefit analysis made by a vendor is seen as callous. Example of Ford? being fined for not fixing a known problem because the cost/benefit was not great enough...

In some cases applications are dependent on the libraries supplied by the vendor of the host platform. One study<sup>853</sup> of Apps running under Android found that those Apps using libraries that contained more reported faults had a slightly smaller average user rating in the Google Play Store.

What motivates developers to fix reported faults in Open Source projects?

- they work for a company who provides software support services for a fee. Having a reputation as the go-to company for a certain bundle of packages is a marketing technique for attracting the attention of anybody looking to spend money on support services, custom modifications to a package or training. Correcting reported faults is a costly signal showing that somebody who knows what they are doing, i.e., status advertising,
- the desire to gain more users may be an incentive for the authors of a system, as a means of enhancing their reputation...
- developers dislike the thought of being wrong or making a mistake; a reported fault preys on their mind and may be fixed to make them feel better, also not responding to known problems in code is not socially acceptable behavior in software development circles. These feelings about what constitutes appropriate behavior are often enough to make developers want to spend time fixing mistakes in code they have written or feel responsible for, provided they have the time. I suspect a lot of problems get fixed by developers when their manager/wife thinks they are working on something more *useful*...

The developer's company is impacted by a fault Google vs Apple... emailed for SQL...?

The cost of fixing a mistake, found prior to release, depends on where it is found in the development cycle... emailed for data...?

The higher the cost of system testing, the fewer opportunities there are likely to be to check systems at this level. Figure 6.7 shows the relationship between the unit cost of a missile (being developed for the US military) and the number of development test flights made.

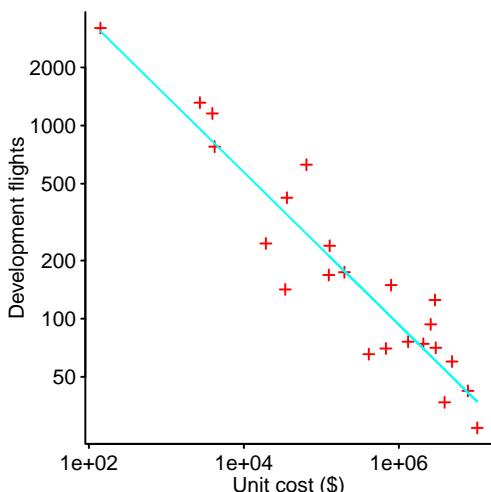


Figure 6.7: Unit cost of a missile against the number of development test flights it had. Data extracted from Augustine.<sup>63</sup> code

## 6.3 Experiencing a fault

Software producing unintended behavior is a consequence of an interaction between a mistake in the code and particular input values.

A program's source code may be riddled with mistakes, but if typical user behavior does not cause the statements containing these mistakes to be executed, the program may gain a reputation for reliability. Similarly, there may only be a few mistakes in the source code, but if they are frequently experienced the program may gain a reputation for being fault-ridden.

Almost all existing research on software reliability has focused on the existence of the mistakes contained in source code. This is convenience sampling, large amounts of Open source is readily available, while information on the characteristics of program input is much harder to obtain.

The greater the number of people using a software system, the greater the volume and variety of inputs it has to process, consequently there are likely to be more faults reported.

A study by Shatnawi<sup>1234</sup> investigated the impact of the number of sites using a release of telecommunication switch software, on the number of software failures reported. A regression model fitted to the data shows that reported faults decreased over time and increase with the number of installed sites (see [reliability/2014-04-13.R](#)).

A study by Lucente<sup>864</sup> investigated help desk incident reports, from 800 applications used by a 100,000 employee company with over 120,000 desktop machines. Figure 6.8 shows...

A comparison of the number of faults reported in different software systems<sup>1092</sup> might be used to estimate the number of people using the different systems; any estimate of system reliability has to take into account the volume of usage and the likely distribution of input values.

Experiencing faults in widely used software is not difficult, e.g., the POSIX library on 15 operating systems<sup>346</sup>...

### 6.3.1 Input profile

The environments in which we live, and software systems operate, often experience regular cycles of activity; events are repeated with small variations, numeric values tend to have values about some mean... common occurrence of the Normal distribution in social science experiments...

Mistakes in code that interact with commonly occurring input values are likely to be noticed and fixed during development; beta testing is a method for discovering common customer usage that developers have not been testing against... some events in everyday life have been found to have an exponential distribution... Anderson library books...

The input profile that resulted in the faults being experienced is an essential aspect of any analysis of program fault characteristics.

Unknown mistakes<sup>ii</sup> exist in shipped systems because the values needed to cause them to experience a fault are not included in the testing process.

Address traces illustrate how the execution characteristics of a program can be dramatically changed by its input. Figure 6.9 shows the number of memory accesses while executing gzip on two different input files. The small colored boxes representing 100,000 executed instruction on the x-axis and successive 4,096 bytes of stack on the y-axis, the colors denote number of accesses within the given block on a logarithmic scale.

A software system processes at least two distinct input profiles, the one used by developers to test it and the one driven by actual usage. Ideally the test and actual usage input profiles are the same, otherwise time and money is wasted fixing mistakes that the customer will not experience. however, even if it is possible to obtain accurate information on the customer input profile, it may be more cost effective to estimate it during testing.

The test process may include automatically generated input cases. Automated techniques include modifying existing tests and various forms of template guided random generation;

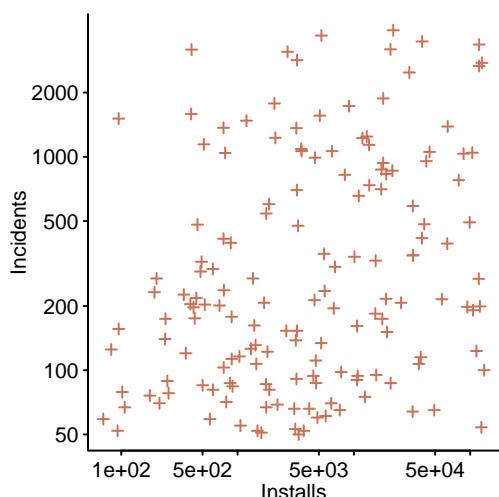


Figure 6.8: Number of reported incidents for each of 800 applications installed on over 120,000 desktop machines. Data from Lucente.<sup>864</sup> [code](#)

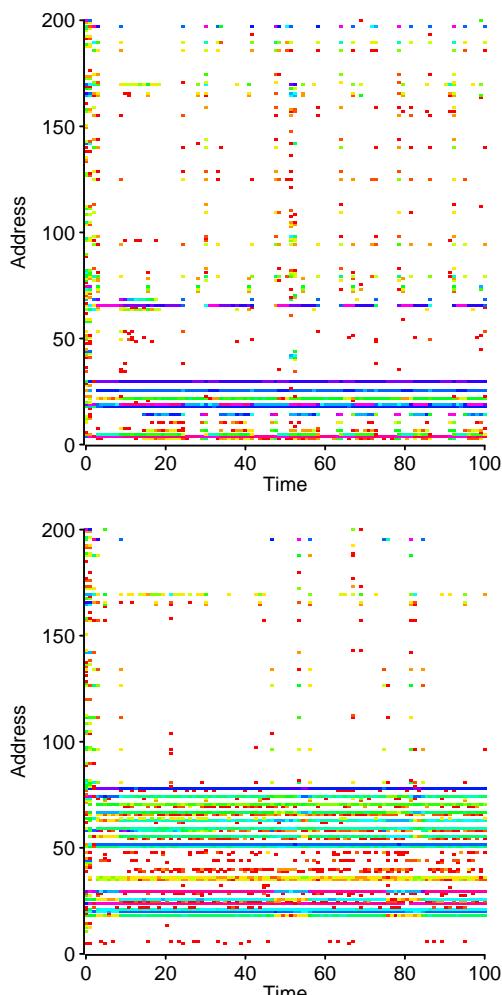


Figure 6.9: Number of accesses to memory address blocks, per 100,000 instructions, executing gzip on two different inputs. Data from Brigham Young<sup>181</sup> via Feitelson. [code](#)

<sup>ii</sup> Management may consider it cost effective to ship a system containing known mistakes.

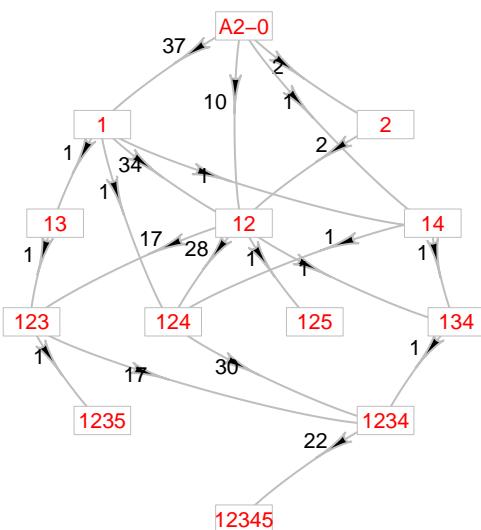


Figure 6.10: Transition counts of five distinct faults experienced in 50 runs of program A2; boxes labeled with the faults experienced up to that point. Data from Nagel et al.<sup>990</sup> code

terms used to denote these processes include *fuzzing* and *mutation*. Fuzzing is often used to test for a particular kind of fault experience, abnormal termination; some tools use a fuzzing selection strategy intended to maximise the likelihood of generating a file that causes a crash fault, e.g., CERT's BFF uses a search strategy that gives greater weight to files which have previously produced faults<sup>642</sup> (i.e., it is a biased random process).

Does the interaction between mistakes in the source code and an input profile result in any recurring patterns in the timing of fault experiences?

One way of answering this question is to count the number of inputs successfully processed by a program between successive fault experiences.

A study by Nagel and Skrivan<sup>990</sup> investigated the timing characteristics of fault experiences in three programs, each written independently by two developers. During execution, each program processed inputs selected from the set of permissible values, when a fault was experienced its identity, execution time up to that point and number of input cases processed were recorded; the coding mistake was corrected and program execution continued until the next fault experience, until five or six faults had been experienced, or the mistake was extremely time-consuming to correct (the maximum number of input cases on any run was 32,808). This cycle was repeated 50 times, always starting with the original, uncorrected, program; the term *repetitive run modeling* was used to denote this form of testing.

Figure 6.10 shows the order in which distinct faults were experienced, by implementation A2 over 50 replications; edge values show the number of times the  $n^{th}$  fault was followed by a particular fault. For example, starting in state A2-0 fault 1 was the first fault discovered during 37 runs and this was followed by fault 3 in one run.

The number of input cases processed before a given number of faults was experienced, during the 50 runs of implementation A2 is shown in the upper plot of Figure 6.11; the lower plot shows the number of inputs processed before each of five distinct faults was experienced.

What is the likely number of inputs that have to be processed by implementation A2 for the sixth distinct fault to be experienced? A regression model could be fitted to the data seen in the upper plot of Figure 6.11, but a model fitted to a sample of five distinct fault experiences will have a wide confidence interval. There is no reason to expect that the sixth fault will be experienced after processing any number of inputs, there appears to be a change point after the fourth fault, but this may be random noise that magnified by the small sample size...

A study by Nagel, Scholz and Skrivan<sup>989</sup> partially replicated and extended the previous study; there were three new developers, two implemented problem three from the first study (in assembler, rather than Fortran) and two implemented problem one from the first study and a new problem...

A study by Dunham and Pierce<sup>366</sup> replicated and extended the work of Nagel and Skriva; problem 1 was independently reimplemented by three developers. The three implementations were each tested with 500,000 input cases, when a fault was experienced the number of inputs processed was recorded, the coding mistake corrected and program execution restarted. This cycle was repeated four times, always starting with the original implementation, fixing and recording as faults were experienced.

Figure 6.12 shows the number of input cases processed, by two of the implementations (only one fault was ever experienced during the execution of the third implementation) before a given number of fault experiences, during each of the four runs. The grey lines are an exponential equation fitted using regression; these two lines show that as the number of faults experienced grows, more input cases are required to experience another fault, and that code written by different developers has different fault rates per input.

A second study by Dunham and Lauterbach<sup>365</sup> used 100 replications for each of the three programs, and found the same pattern of results seen in the first study.

Some published fault experience experiments have used time (computer or user) as a proxy for the quantity of input data. It is not always possible to measure the quantity of input processed and time is often readily available.

A study by Wood<sup>1463</sup> analysed faults found by the product Q/A group in four releases of a subset of products. Figure 6.13 shows that the fault experience rate is similar for the first three releases (the collection of test effort data for release 4 is known to have been different from the previous releases)...

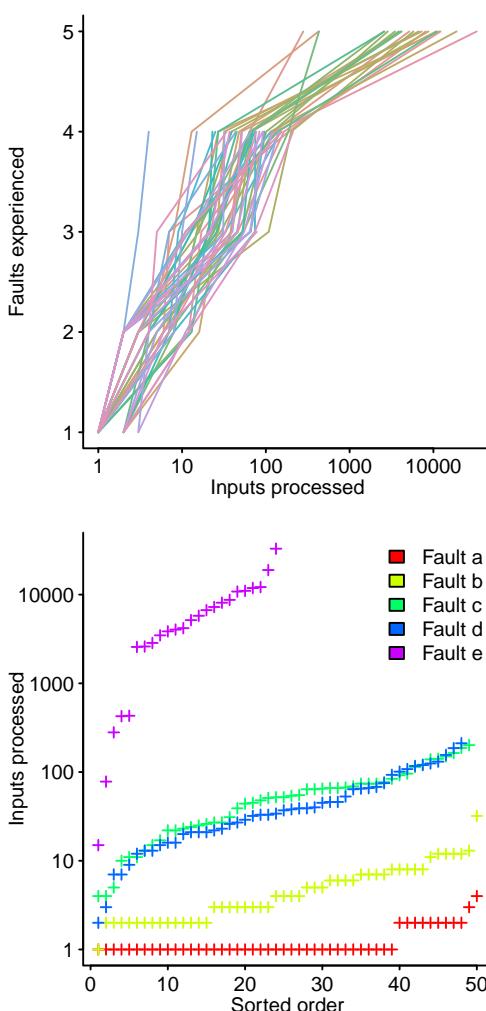
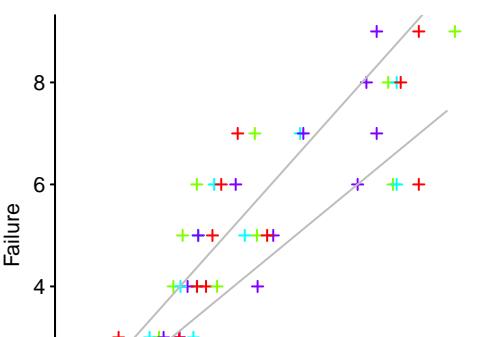


Figure 6.11: Number of input cases processed before a particular fault was experienced by program A2; the list is sorted for each distinct fault. Data from Nagel et al.<sup>990</sup> code



A study by Pradel<sup>1113</sup> searched for thread safety violations, by 15 classes in the Java standard library and JFreeChart declared to be thread safe, and 8 classes in Joda-Time not declared to be thread safe; automatically generated test cases were used. Thread safety violations were found in 22 out of the 23 classes; for each case the testing process was run 10 times and the elapsed time to discover the violation recorded (see Figure 6.14).

A study by Adams<sup>5</sup> investigated faults reported in applications over time. Figure 6.15 shows that approximately one third of all detected faults occurred on average every 5,000 years of execution time (over all users). Only around 2% of faults occurred every five years of execution time.

Multiple occurrences of the same fault provide information about the likelihood that the coding mistake will be executed and about the likelihood of encountering input that can trigger a fault. A regression model using a biexponential equation (e.g.,  $a \times e^{b \times x} + c \times e^{d \times x}$ , where  $x$  is the rank order of occurrences of each fault) has been fitted to the following crash fault data (also see Figure 11.51).

Why is a biexponential model such a good fit? A speculative idea is that the two exponentials are driven by the two independent processes involved in creating the data: the users writing source code and the mistakes contained in the compiler. The users just happen to regularly write particular patterns of code and certain paths through the compiler are more likely than others; each of these two patterns of behavior drives a process that can be modeled by a single exponential.<sup>iii</sup>

A study by Zhao and Liu<sup>1491</sup> investigated the crash faults found by fuzzing the files processed by six Open source programs. Figure 6.16 shows the number of times unique crash faults were experienced in convert and autotrace (estimated by tracing back to a program location), along with lines fitted using biexponential regression models

The reliability of a system depends on the reliability of the components used to process the input. A rarely used component with poor reliability could have less impact on overall system reliability than a frequently used component with high reliability. See the analysis of gcc reliability for an example of using Markov chains to build a model of the reliability of a system based on its components...

Error propagation matrix...?

The consequences of fault experiences not being independent...<sup>207</sup> Error rate varies between programs performing the same function... Correlated failures...

### 6.3.2 Further fault experiences—closed population

This section covers closed populations, i.e., no mistakes are added or removed; it also assumes that the characteristics of the input distribution remain constant.

After  $N$  distinct faults have been experienced, what is the probability that there exists new, previously unexperienced, faults?

Data on reported faults commonly takes two forms: incidence data (e.g., a record of the date of first report and no information on subsequent reports involving the same fault experience), and abundance data (e.g., a record of every fault experience).

It is not possible to use incidence data to distinguish between different exponential order fault growth models<sup>949</sup> (which is nearly all models that have been proposed over the years as potentially applying to a software system). Modeling using incidence data requires samples from multiple sites (e.g., faults experienced within different companies, who use the software, tagged with location-id for each fault experience). It is often possible to fit a variety of equations to fault report data, using regression modeling; however, predictions about future fault experiences made using these models is likely to be very unreliable (see Figure 11.48).

When abundance data is available, the models discussed in Section 4.9.1 can be used to estimate the number of unique items within a population, and the number of new unique items likely to be encountered with additional sampling.

A study by Kaminsky, Eddington and Cecchetti<sup>727</sup> investigated crash faults in three releases of Office and OpenOffice (plus other common document processors), using fuzzing. Figure 6.17 shows a prediction of the growth in crash faults found in the 2003, 2007 and 2010

<sup>iii</sup> Working out which process corresponds to which exponential appearing in the plots is left as an exercise to the reader (because your author has no idea).

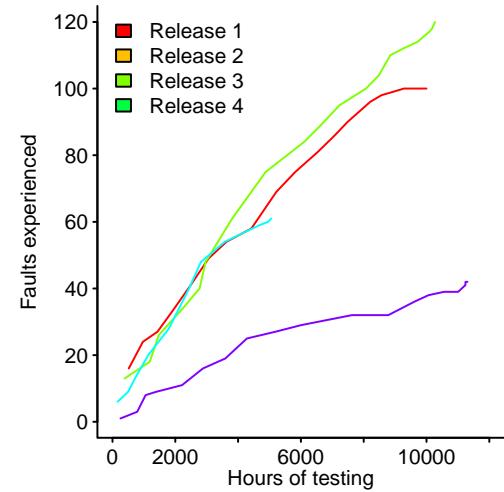


Figure 6.13: Fault discovered against hours of testing, for four releases of a product. Data from Wood.<sup>1463</sup> [code](#)

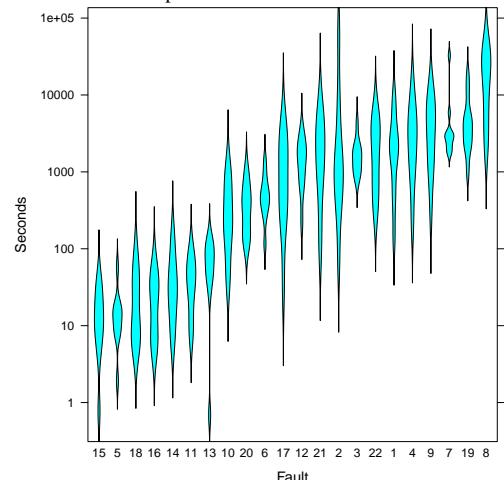


Figure 6.14: Time taken to discover a thread safety violation in 22 Java classes, violin plots for 10 runs involving each class. Data kindly supplied by Pradel.<sup>1113</sup> [code](#)

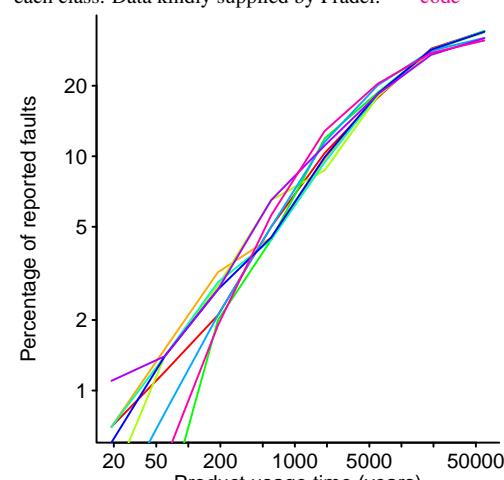


Figure 6.15: Percentage of reported problems having a given mean time to first problem occurrence (in months, summed over all installations of a product) for none products. Data from Adams.<sup>5</sup> [code](#)

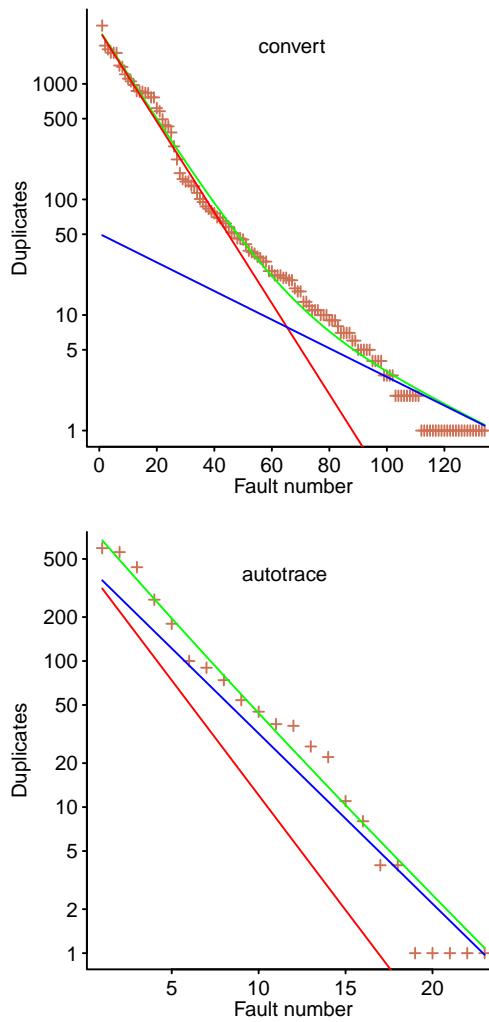


Figure 6.16: Number of times the same fault was experienced in two programs, crashes traced to the same program location; with fitted biexponential equation. Data kindly provided by Zhao.<sup>1491</sup> [code](#)

releases of Microsoft Office, along with 95% confidence intervals. Later versions are estimated to contain fewer crash faults, although the confidence interval for the 2010 release becomes rather wide.

Figure 6.18 shows the number of duplicate crashes experienced when the same fuzzed files were processed by the 2003, 2007 and 2010 releases of Microsoft Office. The blue/purple lines are the two parts of fitted biexponential models...

The previous analysis is based on information about faults that have been experienced. What is the likelihood of a fault experience, given that no faults have been experienced in the immediately previous time,  $T$ ?

An analysis by Bishop and Bloomfield<sup>144</sup> derives a lower bound for the reliability function,  $R$ , for a program executing without failure for time  $t$  after it has executed for time  $T$  without failure; it is assumed that the input profile does not change during time  $T + t$ . The reliability function is:

$$R(t|T) \geq 1 - \frac{t}{T+t} e^{-\frac{T}{t} \log(1+\frac{t}{T})}$$

If  $t$  is much smaller than  $T$ , this equation can be simplified to:

$$R(t|T) \geq 1 - \frac{t}{(T+t) \times e}$$

For instance, if a program is required to execute for 10 hours with reliability 0.9999, the initial failure free period, in hours, is:

$$0.9999 \geq 1 - \frac{10}{(T+10) \times e}$$

$$T \geq \frac{10}{(1-0.9999) \times e} - 10 \approx 36,778$$

If  $T$  is much smaller than  $t$ , the general solution can be simplified to:

$$R(t|T) \geq \frac{T}{t}$$

How can this worst case analysis be improved on?

If it is assumed that a system executes  $N$  times without a failure, and has a fixed probability of failing,  $p$ , the probability of one or more failures occurring in  $N$  executions is (and then assuming  $N$  is large):

$$C = \sum_{n=1}^N p(1-p)^{n-1} \approx p \frac{1-(1-p)^N}{1-(1-p)}$$

How many executions, without failure, need to occur to have a given confidence that the actual failure rate is below a specified level? Rearranging, we get:

$$N = \left\lceil \frac{\log(1-C)}{\log(1-p)} \right\rceil$$

Plugging in values for confidence,  $C = 0.99$ , and failure probability,  $p < 10^{-4}$ , then the system has to execute without failure for 46,050 consecutive runs.

This analysis is not realistic because it assumes that the probability of failure,  $p$ , remains constant for all input cases; studies show that  $p$  can vary by several orders of magnitude.

Multiples reports of the same vulnerability...<sup>606</sup>

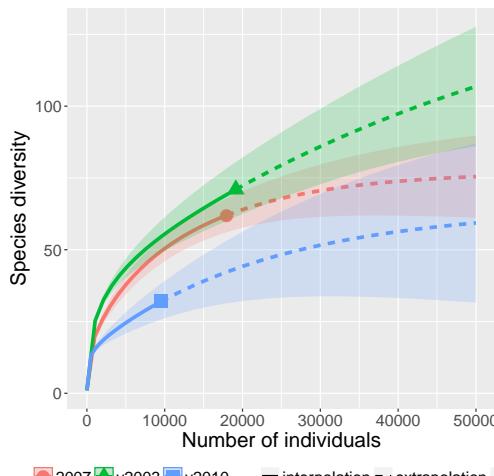


Figure 6.17: Predicted growth, with 95% confidence intervals, in the number of new crash faults found in the 2003, 2007 and 2010 releases of Microsoft Office. Data from Kaminsky et al.<sup>727</sup> [code](#)

■ 2003  
■ 2007  
■ 2010

In an evolving system, existing coding mistakes are corrected and new ones are made; new features may be added that interact with existing functionality, i.e., there may be a change of behavior for the same input. The population of mistakes is open and the input distribution may be evolving.

Bug trackers for Open source projects often contain multiple reports for the same underlying coding mistake, but the population contained within bug tracking systems may be changing because the software system is changing, the user base is changing (i.e., the number of users running a particular version changes as people migrate to a newer release), or both.

Techniques for estimating the characteristics of open populations include: survival analysis of members, ...

How well do sums of exponentials fit the data?

A study by Sun, Le, Zhang and Su<sup>1324</sup> investigated the bugs reported in GCC and LLVM. Figure 6.19 shows the number of times distinct faults reported in GCC (from 1999 to 2015), with a fitted biexponential and the two component exponentials.

A study by Sadat, Bener and Miranskyy<sup>1191</sup> investigated duplicate bug reports in Apache, Eclipse and KDE over 18-years. Figure 6.20 shows the number of times distinct faults reported in KDE, with a fitted triexponential and the three component exponentials. The third component exponential...

Successive releases of a software system often include large amounts of code from earlier releases. The collection of source code that is new in each release can be treated as a distinct ecosystem containing a closed population of mistakes; these ecosystems do not grow but can shrink (when code is deleted). All the code contained in the first release is the foundation ecosystem.

The number of faults that could be experienced in a version of a software system is the sum of the estimated fault experiences that might be triggered by the mistakes in code it contains from the current and earlier releases.

A study by Ozment and Schechter<sup>1045</sup> investigated reported faults in 15 successive versions of OpenBSD, the first in May 1998. The version history of the source was extracted for each release and 140 vulnerabilities (from NVD and several other sources) traced back to the source/version containing the coding mistake.

Table 6.2 shows... not enough data points...

	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7
2.3	5														
2.4	11														
2.5	6		1												
2.6	5	1													
2.7	12	4	2	2	2										
2.8	12	1		1	2										
2.9	4			2											
3.0	3	1			1		2								
3.1	8	2	1	2				1	1						
3.2	6	2						1	2						
3.3	2	1		2							2				
3.4	2				1		1		1						
3.5	7	1	1					2		1					
3.6	3		1												
3.7	1	1			1										

Table 6.2: Reported vulnerabilities in a sequence of versions of OpenBSD (coding mistake made in column version, fault reported in row version). Data from Ozment et al.<sup>1045</sup>

A study by Massacci, Neuhaus and Nguyen<sup>903</sup> investigated 899 Security Advisories in Firefox, reported against six major releases. Their raw data is only available under an agreement that does not permit your author to directly distribute it to readers; the data used in the following analysis was reverse engineered from the paper or extracted by your author from other sources.

The following analysis attempts to build a model of the relationship between the age of code, end-user source code usage and reported faults.

Table 6.3 shows the lowest version (columns) of Firefox containing the mistake in the source code and the highest version (rows) to which a corresponding fault report applies. For instance, 42 faults were discovered in version 2.0 that relate to mistakes assumed to have been made in source code written for version 1.0. Only corrected coding mistakes have been counted, unfixed mistakes are not included in the analysis. Firefox versions have a release and retirement date after which the version is no longer supported (i.e., no more coding mistakes corrected in that version).

How many users does each version of Firefox have at any time?

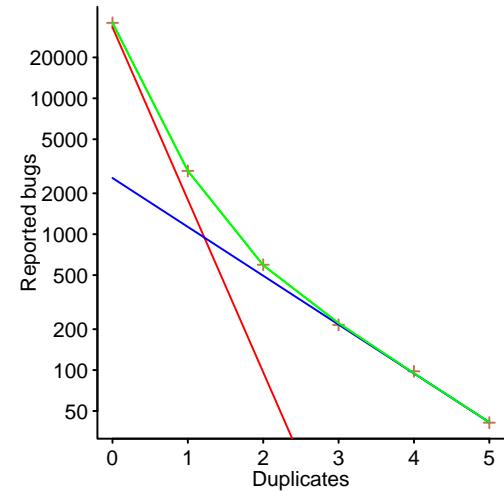


Figure 6.19: Number of instances of the same reported fault in GCC, with fitted biexponential regression model. Data from Sun et al.<sup>1324</sup> code

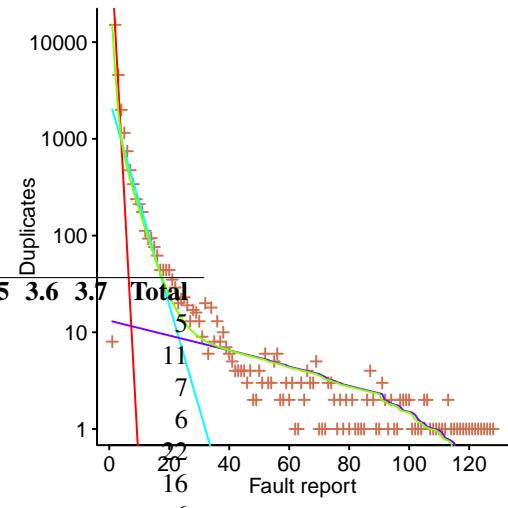
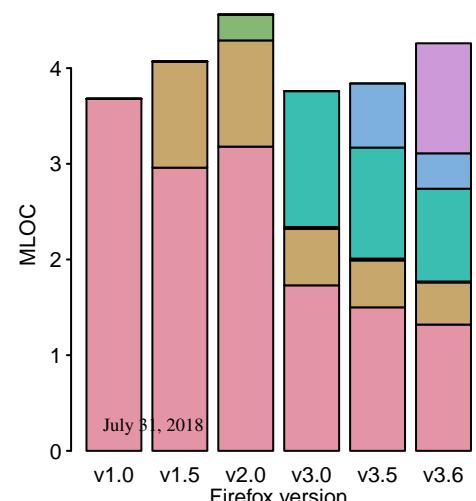


Figure 6.20: Number of instances of the same reported fault in KDE, with fitted triexponential regression model. Data from Sadat et al.<sup>1191</sup> code



	1.0	1.5	2.0	3.0	3.5	3.6
1.0	79					
1.5	71	108				
2.0	42	104	126			
3.0	97	15	22	67		
3.5	32			30	32	
3.6	13		1	5	41	

Table 6.3: Number of reported security advisories in versions of Firefox; coding mistake made in version columns, advisory reported in version row. Data from Massacci et al.<sup>903</sup>

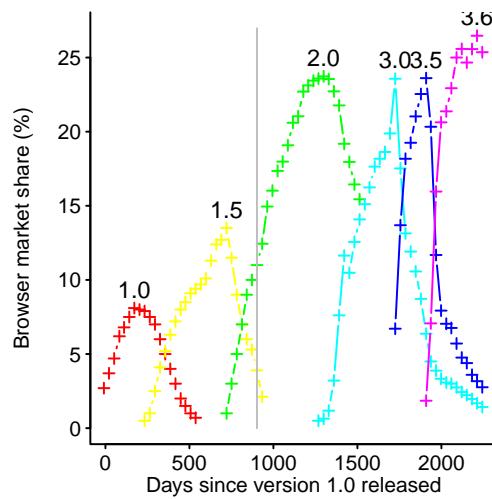


Figure 6.22: Market share of Firefox versions between official release and end-of-support. Data from Jones.<sup>2</sup> [code](#)

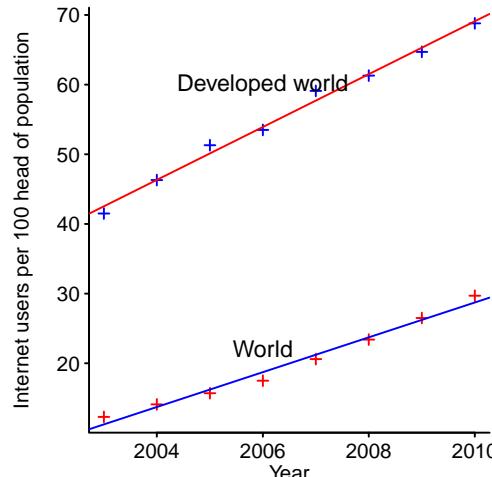


Figure 6.23: Number of people with Internet access per 100 head of population in the developed world and the whole world. Data from ITU.<sup>674</sup> [code](#)

Figure 6.22 shows the market share of the six versions of Firefox between official release and end-of-support. Estimated values appear to the left of the vertical grey line, values from measurements to the right; at its end-of-support date version 2.0 still had a significant market share.

Assuming everybody who uses the Internet uses a browser; Figure 6.23 shows the growth of internet usage over time, broken down by nation development status.

The end-user usage of source code originally written for a particular version of Firefox, over time, is calculated as follows:

- number of lines of code originally written for a particular version contained within the code used to build a later version, or that particular version; call this the build version,
- multiplied by the market share of the build version,
- multiplied by the number of Internet users (the developed world count is used).

Figure 6.24 is an example using the source code originally written for Firefox version 1.0. The green points are the code usage for version 1.0 code executing in Firefox build version 1.0, the orange points the code usage for version 1.0 code executing in build version 1.5 and so on to the yellow points which is the code usage for version 1.0 code executing in build version 3.6. The black points are the sum over all build versions.

Much of the overall growth comes from growth in Internet usage, and in the early years there is also substantial growth in browser market share.

This analysis ignores the possibility that the browsing habits of people who started using the Internet in 2004 may be different from those who first started in 2010, such as time on spent on the Internet, the propensity to report a fault and cultural differences (e.g., European users vs. Chinese users). The contents of web pages also changed...

Combining information on the age of code, end-user source code usage and reported faults, we get...

## 6.4 Where is the mistake?

Information on where mistakes are made during the development of software systems can be used to focus resources on the major problem areas. A few studies<sup>304</sup> have measured across top-level entities such as project phase (e.g., requirements, coding, testing, documentation), while others have focused on specific components (e.g., source code, configuration file), or low level constructs (e.g., floating-point<sup>347</sup>).

The root cause of a mistake, by a person, may be knowledge based (e.g., incorrect beliefs about the semantics of the programming language being used) or it may be skill based (e.g., an error was made in the implementation...)<sup>1148</sup>

Mistakes in hardware<sup>242, 358, 664</sup> tend to occur much less frequently than mistakes in software, and mistakes in hardware are not considered here<sup>iv</sup>

<sup>iv</sup> Your author once worked on a compiler for a cpu that was still in alpha release; the generated code was processed by a sed script to handle known problems in the implementation of the instruction set, problems which changed over the weeks as updated versions of the cpu became available.

The *user interface*, the interaction between people and an application, can be a source of fault experiences in the sense that a user misinterprets correct output, or selects a parameter option that causes program behavior that was not intended.

A study by Nielsen and Landauer<sup>1008</sup> investigated how the number of different usability problems discovered as the number of test subjects increased, using data from 12 studies. Figure 6.25 shows how the number of perceived usability problems increased as the number of test subjects increased (i.e., different people find different things not to like); lines show the fitted regression model  $N(1 - (1 - p)^S)$ , where:  $N$  is the total number of problems available to be found,  $S$  the number of test subjects and  $(1 - p)$  can be interpreted as the probability that a problem will not be found by a subject (both  $N$  and  $p$  are constants returned by the fitting process).

The mistake that causes a fault experience may be in the environment in which a program is built and executed. Many library package managers support the installation of new packages via a command line tool. One study<sup>1366</sup> made use of typos in the information given to command line package managers to cause a package other than the one intended to be installed.

emailed...?

emailed...??

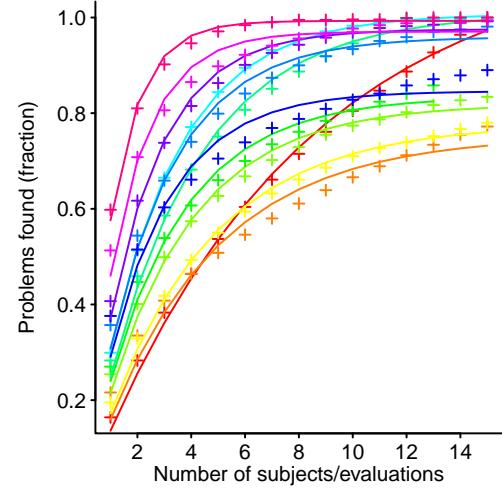


Figure 6.25: Fraction of usability problems found by a given number of test subjects/evaluations in 12 system evaluations, lines show fitted regression model for each system. Data extracted from Nielsen et al.<sup>1008</sup> code

## 6.4.1 Human variability—the random walk of life

Software systems are implemented by generating and interpreting language (human and programming). Reliability is affected by the people's variability in their use of language,<sup>137</sup> what they consider to be correct English syntax<sup>1290</sup> and the interpretation of numeric phrases. Language issues are covered in Section 6.4.3, on source code.

People sometimes have an expectation about what the behavior should be. For instance, people have been found to expect random sequences to have certain attributes,<sup>35,406</sup> e.g., frequent alternation between different values, which from a mathematical perspective represent regularity.

This section discusses numerical quantities that involve uncertainty, e.g., hedge words and qualified expressions, and how people might interpret the chosen value.

Measurements of number usage, in general spoken and written form, show that people prefer to use certain values, either singly (sometimes known as *round numbers*) or as number pairs.

Number pairs (e.g., “10 to 15 hours ago”) have been found to follow a small set of rules,<sup>395</sup> including: the second number is larger than the first, the difference between the values is a divisor of the second value and the difference is at least 5% of the second value.

A round number is any number commonly used to communicate an approximation of nearby values; round numbers are often powers of ten, divisible by two or five, and other pragmatic factors.<sup>684</sup> Round numbers can act as goals<sup>1102</sup> and as clustering points<sup>1280</sup> ...

If a speaker uses a round number, *round\_value\_R*, the probability that the speaker rounded a nearby value to obtain it, is given by:<sup>103</sup>

$$P(\text{Speaker\_rounded}|\text{round\_value\_R}) = \frac{k}{k + \frac{1}{x} - 1}$$

where:  $k$  is the number of values likely to be rounded to *round\_value\_R*, and  $x$  the probability that the speaker chooses to round. Figure 6.26 shows how the likelihood of rounding being used increases rapidly as the number of possible rounded values increases.

An analysis shows that selecting a rounded interpretation yields the greater benefit when there is a small chance that a rounded, rather than non-rounded, use occurred.<sup>103</sup>

Figure 6.27 shows the number of change requests taking a given amount of time to decide whether the change would be made and the time to design+implement those made. There are peaks at the round-numbers 0.5, 1 and 2 hours, with 4 hours perhaps being Parkinson's law target of a half day.

An excess of round numbers has been used to suggest that data has been fabricated<sup>776</sup> ...

People may denote approximate values using numerical expressions containing comparative and superlative qualifiers, such as “more than  $n$ ” and “at least  $n$ ”.

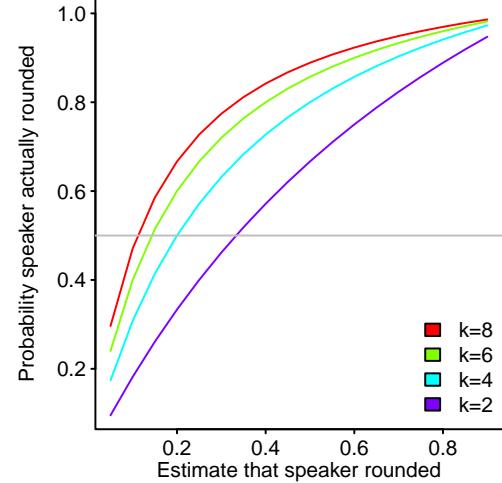


Figure 6.26: Probability the rounded value given has actually been rounded, given an estimate of the likelihood of rounding and the number of values likely to have been rounded; grey line shows 50% probability of rounding. code

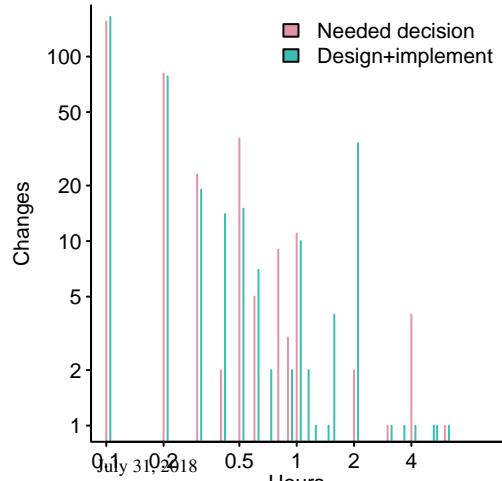


Figure 6.27: Number of change requests having a given recorded time to decide whether needed and to implement

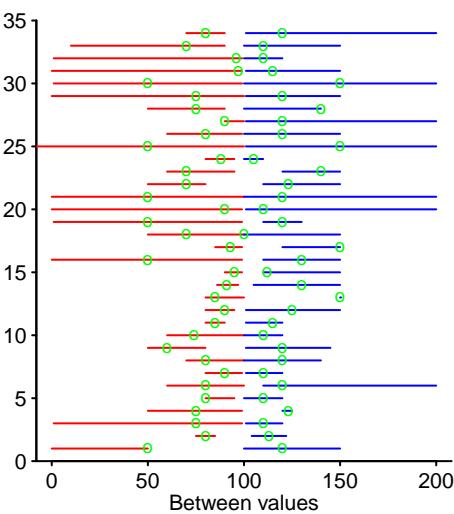


Figure 6.28: Min/max range of values (red/blue lines) and best value estimate (green circles), given by subjects interpreting the value likely expressed by statements containing “less than 100” and “more than 100”. Data kindly provided by Cummins.<sup>302</sup> [code](#)

A study by Cummins<sup>302</sup> investigated the impact of number granularity on the range of values subjects’ assumed it might have, in various kinds of numerical expressions; three numeric granularities were used: *coarse*, e.g., a multiple of 100, *medium* e.g., multiple of 10 and non-multiple of 100, and *fine* e.g., non-round such as 77. Subjects saw statements of the form “So far, we’ve sold fewer than 60 tickets.” (in other statements fewer was replaced by more) and were asked: “How many tickets have been sold? From ?? to ??, most likely ??.”

The results found that the *most likely* value, given by subjects, was closest to the value appearing in the statement when it had a *fine* granularity and furthest away when it was *coarse*; see [reliability/CumminsModifiedNumeral.R](#). Figure 6.28 shows the “From to” range given by each subject, along with their best estimate (in green) for statements specifying “less than 100” and “more than 100”.

A study by Ferson, O’Rawe, Antonenko, Siegrist, Mickley, Luhmann, Sentz and Finkel<sup>429</sup> investigated the language of numerical uncertainty, so called *hedge words*. . . . see [regression/hedges\\_Data.R](#).

Decision by sampling<sup>1314</sup> . . .

People express a preference for liking certain values... [developers/like-n-dis.R](#)  
emailed...?

## 6.4.2 Requirements

A requirements mistake is made when one or more requirements are incorrect, inconsistent or incomplete; an ambiguous specification<sup>135</sup> is a potential mistake. The number of mistakes contained in requirements may be of the same order of magnitude<sup>304</sup> or exceed the number of mistakes found in the code;<sup>1138</sup> different people bring different perspectives to requirements analysis which can result in them discovering different problems.<sup>821</sup>

The same situation can be approached from multiple viewpoints, depending on the role of the viewer; see Figure 2.28. Those implementing a system may fail to fully appreciate all the requirements implied by the specification they are given; context is very important, see Figure 2.27.

The impact of failing to correctly implement a requirement include: going unnoticed, being embarrassing and causing loss of life.

During the lifetime of a project, existing requirements are misinterpreted or changed, and new requirements are added.

Figure 8.27 shows that many mistakes are corrected by modifying the requirements. In those cases where various kinds of behavior are equally acceptable, modifying the requirements documents may be the cheapest path to resolving a mistake...

A study by van der Meulen, Bishop and Revilla<sup>1390</sup> investigated the coding mistakes in 29 thousand implementations of the  $3n + 1$  problem; the programs had been submitted to a programming contest. All submitted implementations were tested and programs producing identical outputs were assigned to the same equivalence class (competitors could make multiple submissions, if the first failed to pass all the tests). In many cases the incorrect output, for an equivalence class, could be explained by a failure of the competitor to implement a requirement implied by the problem being solved, e.g., failing to swap input number pairs when the first was larger than the second.

Figure 6.29 shows the 36 equivalence classes containing the most members; the most common is the correct output, followed by always returning 0 (zero).

The publicly available studies of requirements errors include a study of the software for the Voyager and Galileo spacecraft<sup>874</sup> and . . .

ISO processes require that the committee responsible for a standard maintain a log of reported defects and the committee’s response. Figure 6.30 shows the growth in various kinds of defects reported against the POSIX standard (ISO 9945). ISO language DRs over time...

There have been very few studies<sup>1244</sup> of the impact of the form of specification on its implementation.

Change impact analysis during maintenance...<sup>321</sup>

A connection should exist between lines of code and requirements in that a line of codes exists because of one or more requirements in the specification (ignoring mistakes that result in dead or unreachable code)...<sup>1462</sup> . . .

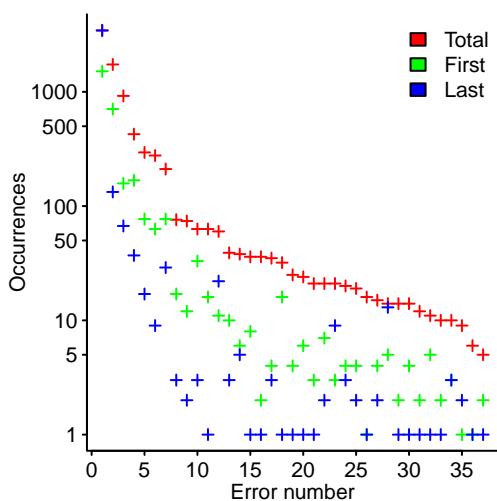


Figure 6.29: Total number of implementations in each of 36 equivalence classes, plus both first and last competitor submissions. Data from van der Meulen et al.<sup>1390</sup> [code](#)

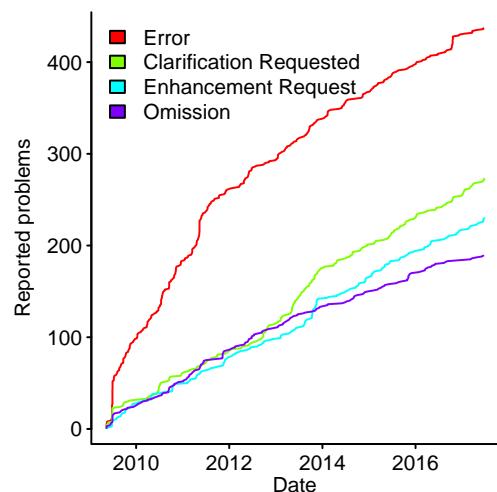


Figure 6.30: Cumulative number of defects logged against the POSIX standard, by defect classification. Data kindly provided by Josey.<sup>1033</sup> [code](#)

### 6.4.3 Source code

Source code is the focus of much research; it is the original form of an executable program that experiences faults and is usually what is modified by developers to correct mistakes. From the research perspective it is now available in bulk, and techniques for analysing it are known and practical to implement.

There are recurring patterns in the changes made to source code to correct mistakes,<sup>1053</sup> one reason for this is that some language constructs are used more often than others.<sup>699</sup> The idea that there is an association between reported faults and particular usage patterns in source code, or program behavior, is popular; over 40 association measures have been proposed.<sup>866</sup>

Errors of omission can cause faults to be experienced. One study<sup>563</sup> of error handling by Linux file systems found that 13% of calls ignore an error code, i.e., do not handle it. Cut-and-paste is a code editing technique that is susceptible to errors of omission, that is, failing to make all the necessary modifications to the pasted version.<sup>122</sup> Significant numbers of cut-and-paste errors have been found in JavaDoc documentation<sup>1041</sup> ...

Patterns of source code usage are also seen in the kinds of mistakes made. Figure 6.31 shows the ranked occurrences of each kind of compiler message generated by Java and Python programs submitted by students.

Proponents of particular languages claim that programs written in the language are more reliable (other desirable characteristics may also be claimed), than if written in other languages. Most experimental studies comparing the reliability of programs written in different languages have either used students<sup>507</sup> or had little statistical power... A language may lack a feature that, if used, would help to reduce the number of mistakes made by developers, e.g., support for function prototypes in C,<sup>1227</sup> which were added to the language in the first ANSI Standard.

Are programs written in some languages more susceptible to experiencing a fault than others? A study by Spinellis, Karakoidas and Lourida<sup>1288</sup> made various kinds of small random changes to 14 different small programs, each implemented in 10 different languages (400 random changes per program/language pair). The ability of these modified programs to compile, execute and produce the same output as the unmodified program was recorded.

Figure 6.32 shows the fraction of programs that compiled, executed and produced correct output, for the various languages. There appear to be two distinct language groupings, each having similar successful compilation rates; one commonality of languages in each group is requiring, or not, variables to be declared before use. The data from the study does not contain enough information to investigate this, or other, possibilities.

The fitted regression model (see [reliability/fuzzer/fuzzer-mod.R](#)) contains an interaction between language and program (the problems implemented did not require many lines of code and in some cases could be solved in a single line in some languages) and a logarithmic dependency on program length (i.e., number of lines).

Another study<sup>104</sup> modified one statement of nine medium-sized Java programs (using abstract syntax trees)... .

Various metrics have been proposed as measures of some desirable, or undesirable, characteristic of a unit of code, e.g., a function. Halstead's and McCabe's cyclomatic complexity are perhaps the most well-known such metrics (see Section 7.2.8, both count the source contained within a single function). Irrespective of whether either of these metrics correlate with anything other than what they directly represent,<sup>808</sup> they can be easily manipulated by splitting functions with high values into two or more functions, which then have lower values of the metric (just as it is possible to reduce the number of lines of code in a function, by putting all the code on one line).

The value of McCabe's complexity (number of decisions, plus one) for the following function is 5, and there are 16 possible paths through the function:

```
int main(void)
{
    if (W) a(); else b();
    if (X) c(); else d();
    if (Y) e(); else f();
    if (Z) g(); else h();
}
```

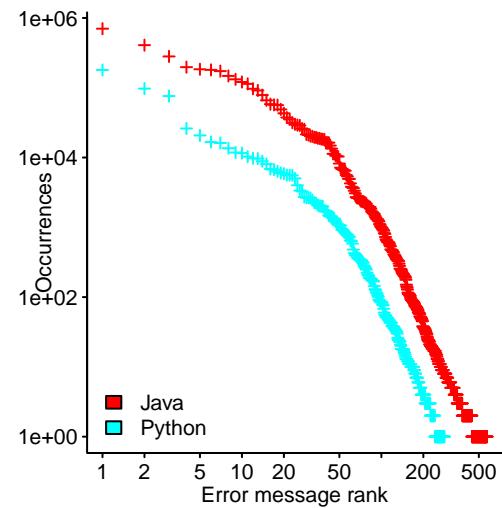


Figure 6.31: Ranked occurrences of compiler messages generated by submitted student Java and Python programs. Data from Pritchard.<sup>1121</sup> [code](#)

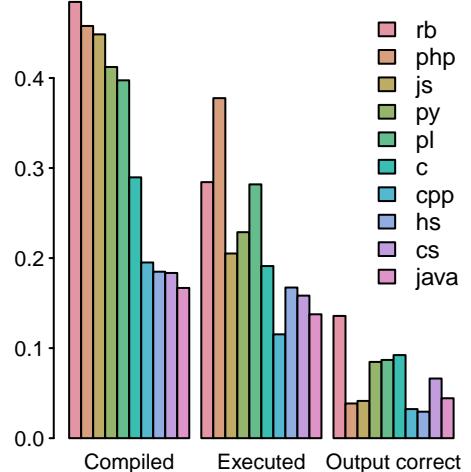


Figure 6.32: Fraction of mutated programs, in various languages, that successfully compiled/executed/produced same output. Data from Spinellis et al.<sup>1288</sup> [code](#)

each `if...else` contains two paths and there are four in series, giving  $2 \times 2 \times 2 \times 2$  paths. Restructuring the code, as below, removes the multiplication of paths caused by the sequence of `if...else`:

```
void a_b(void)
    {if (W) a(); else b();}
void c_d(void)
    {if (X) c(); else d();}
void e_f(void)
    {if (Y) e(); else f();}
void g_h(void)
    {if (Z) g(); else h();}

int main(void)
{
a_b();
c_d();
e_f();
g_h();
}
```

reducing mains McCabe complexity to 1 and the four new functions each have a McCabe complexity of two. Where has the complexity that `main` used to have gone? It now 'exists' in the relationship between the functions, a relationship that is not included in the McCabe complexity calculation. The number of paths that can be traversed, by a call to `main`, has not changed.

A metric that specifies a value for single functions (i.e., its value is calculated from the contents of individual functions) cannot be specified as a control mechanism (i.e., require that values not exceed some limit) because its value can be easily manipulated by moving contents into newly created functions. the software equivalent of what is known as *accounting fraud* in accounting.

Predictions are sometimes attempted<sup>1247, 1500</sup> at the file level of granularity, e.g., predicting which files are more likely to contain mistakes that become fault experiences; the idea being that the contents of some files be rewritten. Any reimplementation will include mistakes and the cost of rewriting the code may be larger than fixing reported faults in the original code as they are discovered.

The idea that there is an optimal value for the number of lines of code in a function body has been an enduring meme (when object oriented programming became popular, the meme mutated to cover optimal class size). See Figure 8.37 for a discussion of the U-shaped defect density paper chase. More reported faults per line of code as the number of lines in a C++ class decreases... see survival analysis<sup>783</sup>...

Fixing most reported faults involves changing a few lines in a single function or changes within a single file. A study<sup>522</sup> of over 1,000 projects for each of C, Java, Python and Haskell found that correcting most coding mistakes involved adding and deleting a few tokens.

A study by Lucia<sup>865</sup> investigated localization techniques. Figure 6.33 shows the percentage of reported faults whose correction involved a given number of files, modules or lines; lines are power laws fitted using regression.

A study by Zhong and Su<sup>1494</sup> investigated commits associated with reported faults in five large Open source projects. Figure 6.34 shows the number of files modified while fixing reported faults, against normalized number of commits made while making these changes; grey line is the equation:  $100 \times \text{files}^{-2.1}$  (which is a reasonable fit to four of the five systems).

Software that is used evolves over time (e.g., lines are added/modified/deleted) and open source programs for which reasonable amounts of data are available are still evolving. See Figure 11.74...

When existing code is changed, there is some probability that a mistake will be made.

A study by Purushothaman and Perry<sup>1124</sup> investigated small source code changes made to one subsystem of the 5ESS telephone switch software (4,550 files containing almost 2MLOC,

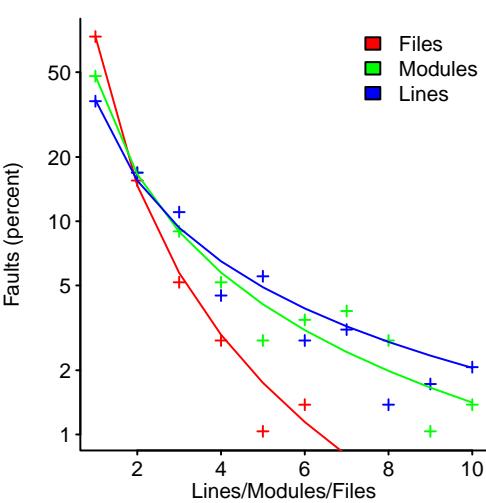


Figure 6.33: Number of reported faults whose fixes involved a given number of files, modules or lines in a sample of 290 faults in AspectJ; lines are power laws fitted using regression. Data from Lucia.<sup>865</sup> code

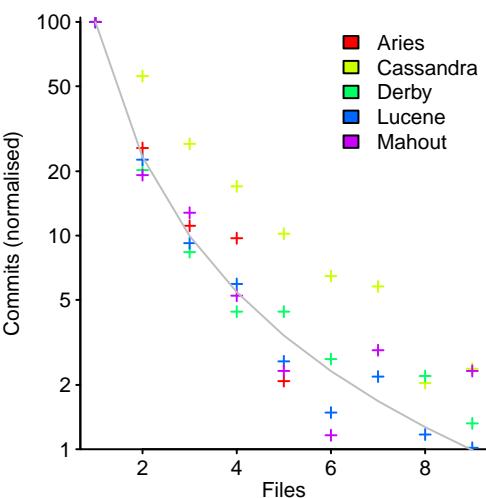


Figure 6.34: Normalized number of commits, made to fix reported faults, involving a given number of files in five software systems; grey line is power law fitted using regression. Data from Zhong et al<sup>1494</sup> via M. Monperrus. code

with 31,884 modification requests after the first release changing 4,293 of these files). Figure 6.35 is based on an analysis of reported faults traced to updates, that involved modifying/inserting a given number of lines, and shows the percentage of each kind of update that resulted in a reported fault.

Some languages have a mechanism for developers to specify properties, about variables appearing in the source, that are expected to be true, e.g., `assert(x > y)...`??

#### 6.4.4 Third-party libraries

The availability of many open source libraries can make it more cost effective to interface to a third-party, rather than implementing a bespoke solution.

A study by Decan, Mens and Constantinou<sup>331</sup> investigated the time taken for security vulnerabilities in npm packages to be fixed, along with the time taken to update packages that depended on a version of a package having a reported vulnerability. Figure 6.36 shows survival curves (with 95% confidence bounds), for high and medium severity vulnerabilities, of time to fix a reported vulnerability (Base) and time to update a dependency (Depend) to a corrected version of a package.

#### 6.4.5 Correctness vs. consistency

Intel trig functions...

Testing the quality of random number generators<sup>824</sup>

Errors in Microsoft Excel's random number generator<sup>916</sup>

#### 6.4.6 Documentation

Documentation is a cost paid today, that may provide a benefit later for somebody else or the author.

If user documentation specifies functionality that is not supported by the software, the vendor may be liable to pay damages to customers expecting to be able to perform the documented functionality.<sup>732</sup> Non-existent documentation is not unreliable, but documentation that has not been updated to match changes to the software is.

There have been relatively few studies of the reliability of documentation. A study by Rubio-Gonzalez and Libit<sup>1185</sup> found 1,784 undocumented error return codes in an analysis of the source code of 53 Linux file systems. A study by Ma, Liu and Forin<sup>877</sup> tested an Intel x86 cpu emulator and found a wide variety of errors in the documentation specifying the behavior of the processor.

### 6.5 Non-software causes of unreliability

Hardware is built from moving parts and wears out. Electronic components operate by controlling the direction of movement of electrons; when the movement is primarily in one direction, atoms migrate in that direction, and over time this migration degrades device operating characteristics.<sup>1293</sup> A factor with ever smaller transistors is their decreasing mean time to failure; expected chip lifetimes have dropped from 10 years to 7, and decreasing.<sup>1443</sup>

As the size of components shrinks and the number of components on a device increases, the probability that thermal noise will cause a bit to change state increases.<sup>762</sup>

Faulty hardware does not always noticeably change the behavior of a program when it is executing, apparently correct program execution can occur, e.g., image processing.<sup>1021</sup>

See Table 13.3 for a discussion of system failure traced to either cpu or DRAM failures and Section 10.6 for a study that looked for a correlation between checking for a hardware performance and likelihood of experiencing some intermittent faults.

A software reliability problem that was rarely encountered outside of science fiction a few decades ago, now regularly occurs in modern computers: cosmic rays (plus more local

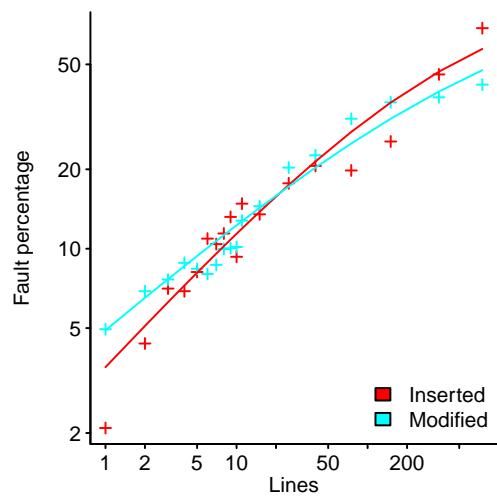


Figure 6.35: Percentage of insertions/modifications of a given number of lines resulting in a reported fault; lines are fitted regression models. Data from Purushothaman et al.<sup>1124</sup> code

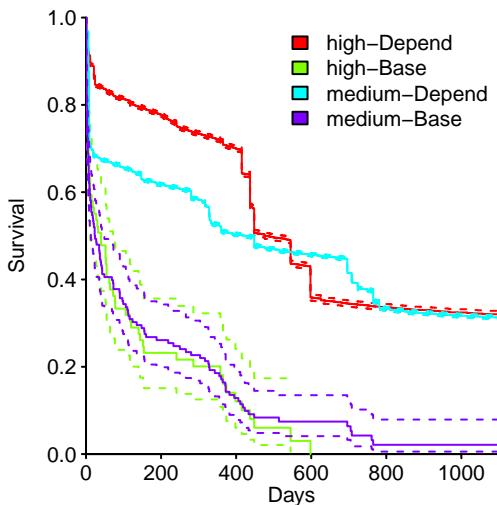


Figure 6.36: Survival curve (with 95% confidence bounds) of time to fix vulnerabilities reported in npm packages (Base) and time to update a package dependency (Depend) to a corrected version (i.e., not containing the reported vulnerability); for vulnerabilities with severity high and medium. Data from Decan et al.<sup>331</sup> code

sources of radiation, such as the materials used to fabricate a device) flipping the value of one or more bits in memory or a running processor. Techniques for mitigating the effects of a radiation induced events have been proposed.<sup>983</sup>

The two main sources of radiation are alpha-particles generated within the material used to fabricate and package devices and Neutrons generated by Cosmic-rays impacting with the upper atmosphere.<sup>66</sup> The data in Figure 6.37 comes from monitoring equipment located in the French Alps; either, 1,700m under the Fréjus mountain (i.e., all radiation is generated by the device), or on top of the Plateau de Bure at an altitude of 2,552m (i.e., radiation sources are local and Cosmic).<sup>65</sup> For confidentiality reasons the data has been scaled by a small constant.

Figure 6.37 shows how the number of bit-flips increased over time (measured in Mega-bits per hour), for SRAM fabricated using 130nm, 65nm and 40nm processes. The 130nm and 65nm measurements were made underground and the lower rate of bit-flips for the 65nm process is the result of improved materials selection, reducing alpha-particle emissions; the 40nm measurements were made on top of the Plateau de Bure and show the impact of external radiation sources.

The soft error rate is usually quoted in FITs (Failure in Time), with 1 FIT corresponding to 1 error per  $10^9$  hours per megabit, or  $10^{-15}$  errors per bit-hour. Consider a system with 4 GB of DRAM (1000 FIT/Mb is a reasonable approximation for commodity memory,<sup>1348</sup> which increases with altitude and is 10 times greater in Denver, Colorado), it has an MTBF of  $1000 \times 10^{-15} \times 4.096 \times 10^9 \times 8 = 3.2 \times 10^{-2}$  hours, around once every 33 hours. Soft errors are a regular occurrence for installations with hundreds of terabytes of memory.<sup>657</sup>

The Cassini spacecraft experienced an average of 280 single bit memory errors per day<sup>1331</sup> (in two identical flight recorders containing 2.5G of DRAM). The rate of double-bit errors was higher than expected (between 1.5 and 4.5%) because the incoming radiation had enough energy to flip more than one bit.

Uncorrected soft errors place a limit on the maximum number of computing nodes that can be usefully used by one application; at around 50,000 nodes, a system would spend half its time saving checkpoints and restarting from previous checkpoints after an error occurs.<sup>1160</sup>

Error correcting memory reduces the probability of an uncorrected error by several orders of magnitude, but with modern systems containing terabytes the probability of an error adversely affecting the result is high.<sup>657</sup> The Cray Blue Waters system at the National Center for Supercomputing Applications experienced 28 uncorrected memory errors (ECC and Chipkill parity hardware checks corrected 722,526 single bit errors and 309,359 two-bit errors, a 99.995% success rate).<sup>348</sup> Studies<sup>931</sup> have investigated assigning variables deemed to be critical to a subset of memory that is protected with error correcting hardware, along with various other techniques.<sup>870</sup>

Calculating the FIT for processors is complicated.<sup>840</sup>

Redundancy can be used to continue operating after experiencing a hardware fault, e.g., three processors performing the same calculating and a majority vote used to decide which of the outputs to accept.<sup>1474</sup> Software only redundancy techniques include having the compiler generate, for each source code sequence, two or more independent machine code sequences<sup>1152</sup> whose computed values are compared at various check points, and replicating computations across multiple cores<sup>1489</sup> (and comparing outputs). The overhead of duplicated execution can be reduced by not replicating those code sequences that are less affected by register bit flips<sup>423</sup> (e.g., the value returned from a bitwise AND that extracts 8 bits from a 32-bit register is 75% less likely to deliver an incorrect result than an operation that depends on all 32 bits). Optimizing for reliability can be traded off against performance,<sup>959</sup> e.g., ordering register usage such that the average interval between load and last usage is reduced.<sup>1469</sup>

Developers don't have to rely on compiler or hardware support, reliability can be improved by using algorithms that are robust in the presence of *faulty* hardware. For instance, the traditional algorithms for two-process mutual exclusion are not fault tolerant; a fault tolerant mutual exclusion algorithm using  $2f + 1$  variables, where a single fault may occur in up to  $f$  variables is available.<sup>970</sup> Researchers are starting to investigate how best to prevent soft errors corrupting the correct behavior of various algorithms.<sup>186</sup> Bombarding a system with radiation increases the likelihood of radiation induced bit-flips.<sup>941</sup>

The evolution of commodity cpu and memory chips appears to be heading towards cheap and unreliable products, just like many household appliances are priced low and have a short expected lifetime...

A study by Dinaburg<sup>355</sup> found occurrences of bit-flips in domain names appearing within HTTP requests, e.g., a page from the domain ikamai.net being requested rather than from akamai.net. The  $2 \cdot 10^{-9}$  bit error rate was thought to occur inside routers and switches. Undetected random hardware errors can be used to redirect an access to another site<sup>355</sup> ...

If all the checksums involved in TCP/IP transmission are enabled, the users visiting Facebook on average once per day and downloading 2M of Javascript per visit, gives an expected bit flip rate of once every 5 days somewhere in the world.

The impact of level of compiler optimization on a program's susceptibility to bitflips is discussed in Section 11.2.2.

### 6.5.1 System availability

A system is only as reliable as the least unreliable critical subsystem, and the hardware on which software runs is a critical subsystem that needs to be included in any application reliability analysis; some applications also require a working internet connection, e.g., for database access.

Before cloud computing became a widely available commercial service, companies built their own clustered computer facilities (low usage rates of such systems<sup>667</sup> is what can make cloud providers more cost effective).

The reliability of Internet access to the services provided by other computers is not yet high enough for people to overlook the possibility that failures can occur<sup>132</sup> (see Section 13.3.6)...

Long-running applications need to be able to recover from hardware failures, if they are to stand a reasonable chance of completing. A process known as *checkpointing* periodically stores the current state of every compute unit, so that when any unit fails, it is possible to restart from the last saved state, rather than restarting from the beginning. A tradeoff has to be made<sup>1472</sup> between frequency of checkpointing, which takes resources away from completing execution of the application but reduces the amount of lost calculation, and infrequent checkpointing, which diverts less resources but incurs greater losses when a fault is experienced. Calculating the optimum checkpoint interval<sup>307</sup> requires knowing the distribution of node uptimes; see Figure 6.38.

The Los Alamos National Laboratory (LANL) has made public, data from 23 different systems installed between 1996 and 2005.<sup>814</sup> these systems run applications that "... perform long periods (often months) of CPU computation, interrupted every few hours by a few minutes of I/O for check-pointing." Figure 6.38 shows the 10-hour binned data fitted to a zero-truncated negative binomial distribution for systems 2 and 18.

Operating systems and many long-running programs sometimes write information about a variety of events to one or more log files. One study<sup>1478</sup> found that around 1 in 30 lines of code in Apache, Postgresql and Squid was logging code; this information was estimated to reduce median diagnosis time by a factor of 1.4 to 3. The information diversity of system event logs tends to increase, with new kinds of information being added, but the writing of older information not being switched off (because it might be useful); log files have been found to contain<sup>1031</sup> large amounts of low value information, more than one entry for the same event, changes caused by software updates, poor or no documentation and inconsistent information structure within entries.

## 6.6 Checking for intended behavior

The two main methods for checking that code behaves as intended, are: analyzing the source code to work out what it does and reviewing the behavior of the code during execution (e.g., testing). Little data is available on the kinds of problems found and the relative cost-effectiveness of the various techniques that are used.

Many software systems support a range of optional constructs and support for these may be selected by build time configuration options. When checking for intended behavior, one of the first questions will involve the versions of the system being checked. Some systems support so many options that simply checking whether all possible configurations can be built requires a huge investment of time and computing resources.<sup>572</sup>

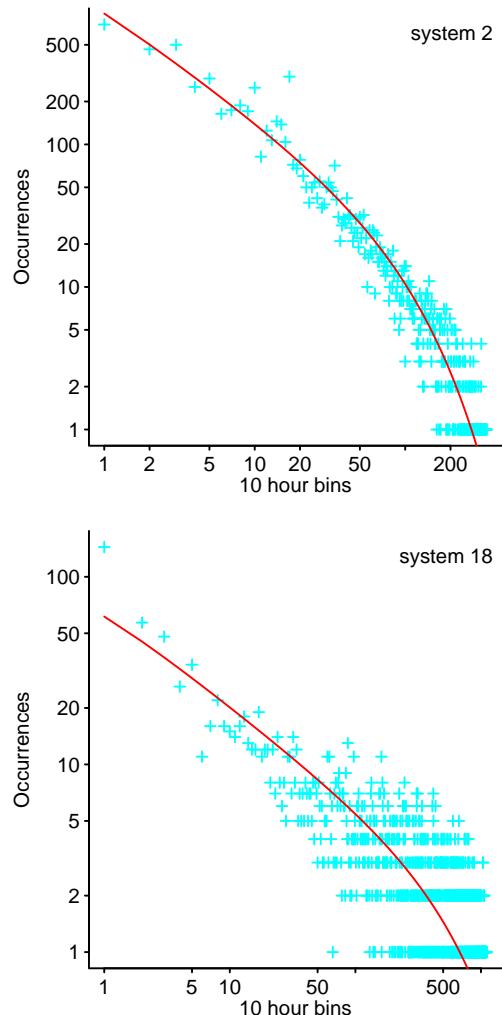


Figure 6.38: For systems 2 and 18, number of uptime intervals, binned into 10 hour intervals, red line is fitted negative binomial distribution. Data from [Los Alamos National Lab \(LANL\). code](#)

The further along in the development process a problem is found, the more costly it is likely to be to correct it (possible additional costs include having to modify something created between the introduction of the problem and its detection, and having to recheck work). This additional cost does not necessarily make it more cost effective to detect problems as early as possible. The relative cost of correcting problems vs. detecting problems may make it more cost effective to check for problems at specific point in the development process.

The cost of correcting problems will depend on the cost characteristics of the system containing the software; developing software for a coffee vending machine is likely to be a lot cheaper than for a jet fighter because of the cost of the hardware needed to testing. Data from NASA and US Department of Defense, on relative costs of fixing problems discovered in various phases of development are very large because of the very high cost of the hardware running the software systems developed for these organizations.

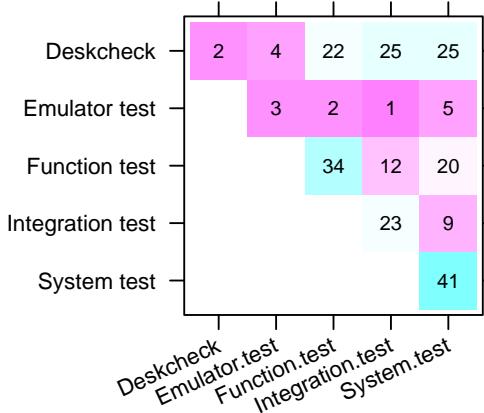


Figure 6.39: Fault slip throughs for a development project at Ericsson; y-axis lists phase when fault could have been detected, x-axis phase when fault was found. Data from Hribar et al.<sup>645</sup> code

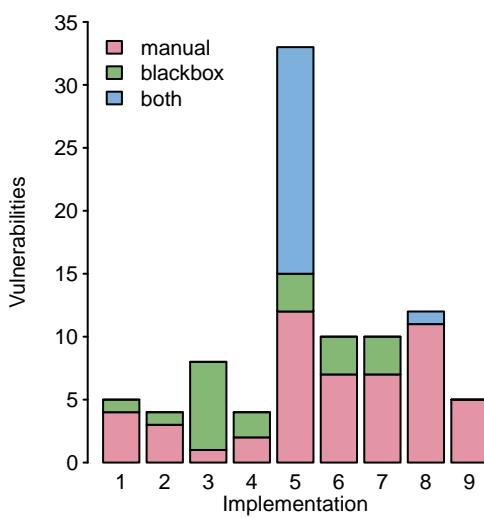


Figure 6.40: Number of vulnerabilities found using blackbox testing and manual code review of nine implementations of the same specification. Data from Finifter.<sup>434</sup> code

To reduce time and costs the checking process may be organized into different levels of granularity, not proceeding to a higher level until the lower levels are known to work: unit testing is performed by individual developers, integration testing checks that multiple components or subsystems work together and systems testing is performed on the system as a whole.

A study by Hribar, Bogovac and Marinčić<sup>645</sup> investigated *Fault Slip Through*, they analysed the development phase where a fault was found compared to where it could have been found. Figure 6.39 shows the number of faults found in various test phases (deskcheck is a form of code review performed by the authors of the code) and where the fault could have been found (as specified on the fault report; also see Antolić<sup>1382</sup>).

The problem experienced by experiments designed to compare different review methods is that the variation in skill and knowledge between the subjects, and variation in the difficulty of finding different kinds of mistakes, is often much greater than performance differences that might be attributed to the different techniques.

A study by Myers<sup>984</sup> investigated the faults experienced using program testing and code walkthroughs/inspections. The 59 professional developers were split into two groups of 16 individual teams, who ran tests (either just with a specification, or with both a specification and program listing), and nine teams of three, who did walkthroughs/inspection... See [reliability/myers1978.R](#) todo.

A study by Finifter<sup>434</sup> investigated the mistakes found and fault experienced using manual code review and black box testing of nine implementations of the same specification. Figure 6.40 shows the number of vulnerabilities found by the two techniques in the nine implementations; some of the difference is due to the variation in the abilities and kinds of mistakes made by different developers, plus skill differences in using the programming languages.

While there has been a flurry of activity applying machine learning techniques to fault prediction, the models have not generalized outside the data used to build them (even between different versions of the same project;<sup>1499</sup> see [faults/eclipse/eclipse-pred.R](#)). Very noisy data is one problem, along with no usage data (see Section 6.1.3).

## 6.6.1 Review meetings

Review meetings support a variety of functions, including: highlighting of important information between project members (i.e., ensuring that people are kept up to date with what others are doing) and uncovering potential problems before changing things becomes much more expensive.

Most research has used problems-found as the metric for evaluating review meetings, in particular potential fault experiences found during code reviews. Problems found is something that is easy to measure, code is readily available and developers to review it are likely to be more numerous than people with the skills needed to review requirements and design documents (which do not always exist, as such).

The range of knowledge and skills needed to review requirements and design documents may mean that some of those involved concentrate on areas that are within their domain of expertise.<sup>378</sup> Many of the techniques used for estimating population size assume that capture sites (i.e., reviews) have equal probabilities of flagging an item, and estimates based on data from meetings where reviewers have selectively read documents will be biased; see [reliability/eickt1992.R](#).

Detecting problems may not even be the main reason for performing code reviews,<sup>73</sup> keeping teams members abreast of developments and creating an environment of shared ownership of code may be considered more important.

People tend to have a repertoire of actions, which are used on a regular basis,<sup>7</sup> and result in mistakes being introduced into the code. Recurring coding patterns associated with known mistakes can be searched for, in code, and flagged. Depending on the construct the search process may best be performed by a tool or by a developer reviewing the code.

Ideally code flagged by a tool is incorrect (e.g., reading from an uninitialized variable), but the analysis performed may not be sophisticated enough to handle all possibilities<sup>1202</sup> (e.g., there is some uncertainty about whether the variable being read from has been written to) or the usage may simply be suspicious (e.g., use of assignment in the conditional expression of an if-statement, when an equality comparison was intended, i.e., = had been typed instead of ==). The issue of how developers might respond to false positive warnings is discussed in Section 9.1.

A study of one tool<sup>1493</sup> found a strong correlation between flagged code and faults experienced during testing and faults reported by customers (after the output of the tool had been cleaned by a company specializing in removing false positive warnings from static analysis tool output).

Traditionally a human code review (other terms include *code inspection* and *walkthroughs*<sup>459</sup>) has involved one or more people reading another developer's code, and then meeting with the developer to discuss what they have found. These days the term is also associated with maintainers reviewing code that has been pushed to a project's version control system, to check that it is ok to merge...

A variety of different code review techniques have been proposed, including: Ad-hoc (no explicit support for reviewers), Checklist (reviewers work from a list of specific questions that are intended to focus attention towards common problems), Scenarios-based (each reviewer takes on a role intended to target a particular class of problems) and Perspective-based reading (reviewers are given more detailed instructions, than are given in Scenario-based reviews, about how to read the document; see Section 13.2 for an analysis)...

A study by Hirao, Ihara, Ueda, Phannachitta and Matsumoto<sup>621</sup> investigated the impact of positive and negative code reviews on patches being merged or abandoned (for Qt and OpenStack). A logistic regression model found that for Qt positive votes were more than twice as influential, on the outcome, as negative votes, while for, OpenStack negative votes were slightly more influential (see [reliability/OSS2016.R](#)).

A study by Porter, Siy, Mockus and Votta<sup>1104</sup> recorded code inspection related data from a commercial project over 18 months (staffed by six dedicated developers and five developers who also worked on other projects). The best fitted regression model had the number of mistakes found being proportional to the log of the number of lines reviewed and the log of meeting duration; this study is discussed in Section 13.2, also see Figure 11.32.

## 6.6.2 Testing

The purpose of testing is to gain some level of confidence that the software operates in a way that is likely to be acceptable to the customer, given how the vendor intends it to perform (e.g., meets acceptance payment criteria or is unlikely to experience a high rate of customer returns).

Most published research of testing has been at the system level...

During testing, one indication that the software is becoming more reliable, is that the number of previously unseen faults experiences per unit of test effort is decreasing; another reason for a decrease in new fault discovery is replacement of existing testers by less skilled staff. The extent to which the reliability, as experienced by the customer, increases depends on the overlap between test inputs and likely customer inputs to the software.

A study by Stikkel<sup>1316</sup> investigated three industrial development projects. Figure 6.41 shows the (normalised) number of faults discovered per man-hour of testing, averaged over a week, for these projects; two show a sharp decline in new fault discoveries, per man-hour, in the closing weeks of testing, while the third shows no such pattern.

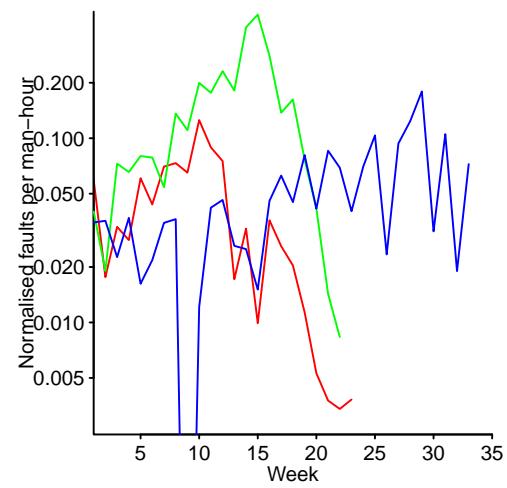


Figure 6.41: Number of faults experienced per unit of testing effort, over a given number of weeks. Data from Stikkel.<sup>1316</sup> [code](#)

The number of possible combinations of inputs to a non-trivial program is so large, that by comparison, the number of tests that are practical to run is insignificant. However, the number of input combinations used by customers may be small enough to be make thorough testing viable, but the detailed information on the input combinations used by customers is often unavailable.

An example of how the input used for random testing can be unrepresentative of customer input is provided by a study<sup>260</sup> that performed random testing of the Eiffel base library. The functions in this library contain extensive pre/post condition checks, and random testing found twice as many mistakes in these checks as the implementation of the functionality; the opposite of the pattern seen in user reported faults.

Sources of manually created tests include developers packaging up tests they have written to check their own code, fault reports submitted by users and in some cases third-party written test suites. Some customers are sufficiently interested in the behavior of the software they buy, they are willing to fund the development of a test suite. The US Government appreciated the economic advantages of being able to select from multiple hardware vendors, and ensuring that compilers from different vendors could compile the same source code was an important factor in increasing program portability; validation suites were funded for Cobol and Fortran,<sup>1032</sup> and later SQL, POSIX<sup>7</sup> and Ada.<sup>4</sup>

Manual tests can be very effective, but creating them is very time-consuming<sup>?,1032</sup> and expensive. Various kinds of automatic test generation are available.

Testing that the behavior of a program is as intended requires knowledge of the intended behavior, for a given input. While some form of automatic input generation is possible, in only a few cases<sup>43</sup> is it possible to automatically predict the expected output from the input, independently of the software being tested. One form of program behavior is easily detected, abnormal termination, and some forms of fuzz testing use of this as their correctness criteria.

When multiple systems supporting the same functionality are available, it may be possible to use differential testing to compare the outputs produced from a given input (a difference being a strong indicator that one of the systems is behaving incorrectly)<sup>244</sup>...

While studies<sup>789</sup> have found that the majority of faults are experienced with test cases involving a change of one input value, this result may be due to most test cases involving a single input value or some other reason (the necessary information is rarely provided)...

ISO Standards have been published covering methods for measuring conformance to particular standards<sup>671,672</sup> and requirements for test laboratories.<sup>670</sup> However, the lack of commercial interest...

Tests may need to be updated when code changes, and new tests written to handle any new modified behavior; maintenance can be a significant expense.<sup>?</sup> Like production code, tests can contain mistakes.<sup>1384</sup>

An analysis of a few projects<sup>1480</sup> found no correlation between the growth of project and test code (see [time-series/argouml\\_complete.R](#)).

To what extent do test suites change over time? A study by Marinescu, Hosek and Cadar<sup>896</sup> measured the statement and branch coverage (using the test suite distributed with the program's source) of six open source programs over time. Figure 6.42 shows that for some widely used programs to statement coverage of the test suite did not vary much over five years.

Evolution of testing library usage...?

### 6.6.2.1 Combinatorial testing

emailed for data...? emailed...?  
?

### 6.6.2.2 Beta testing

The people who use software systems may operate it using input profiles that are very different from those envisaged by the developers who tested the software. Beta testing is a way

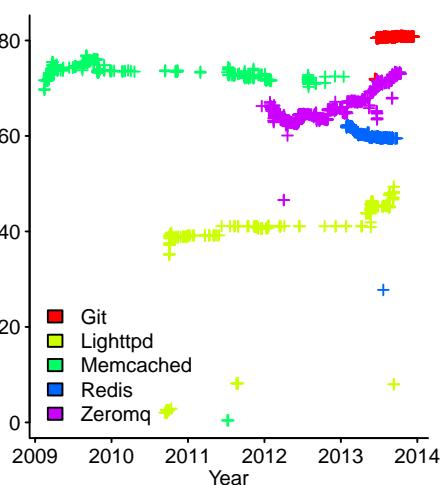


Figure 6.42: Statement coverage achieved by the respective program's test suite (data on sixth program not usable). Data from Marinescu et al.<sup>896</sup> [code](#)

of discovering problems with software (e.g., coding mistakes and incomplete requirements), when processing the input profiles of the intended users. The number of problems found during beta testing, compared to internal testing provides feedback on the relevance of the usage profile that drives the test process.<sup>892</sup>

Behavior during beta testing different from after release... Figure are images :-?

### 6.6.2.3 Estimating test effectiveness

Is the test process for a software system able to reliably provide the desired level of confidence that the software is good enough to be acceptable to the customer?

A necessary requirement for checking the behavior of code is to execute it, every statement not executed is untested. Various coverage criteria are available, for instance percentage of program statements or branches executed by the test process (the coverage achieved by a test suite is likely to vary between platforms and even compiler used<sup>476</sup> more???).

A study by Inozemtseva and Holmes<sup>663</sup> investigated test coverage of five very large Java programs. The results showed a consistent relationship between percentage statement coverage,  $sc$ , and percentage branch coverage,  $bc$  (i.e.,  $bc \propto sc^{1.2}$ ), and percentage modified condition coverage,  $mc$  (i.e.,  $mc \propto sc^{1.7}$ ); see [reliability/coverage\\_icse-2014.R...](#)

Software written for safety critical applications often has to have 100% MC/DC coverage<sup>251</sup>... no data found...

A study by Gopinath, Jensen and Groce<sup>521</sup> investigated some of the characteristics of coverage metrics for the test suites of 1023 Java projects. Figure 6.43 shows fraction of statement coverage against branch coverage; each circle is data from one project. The lines are fitted regression models, which contain a strong interaction between coverage and log  $KLOC$ ... no correlation between  $\log(loc)$  and test suite coverage...

A study by Kang, Ray and Jana<sup>734</sup> investigated the number of statements contained along the execution paths, within a function, followed after a call to a function that could return an error, i.e., the error and non-error paths. Figure 6.44 shows that non-error paths often contained more statements than the error paths...

One explanation for the slow initial growth in branch coverage seen in Figure 6.43 is that the test suites primarily contain positive tests (which contain more statements)...

A study by Čaušević, Shukla, Punnekkat and Sundmark<sup>1375</sup> found that developers wrote almost twice as many positive tests as negative tests, for the problem studied; see [reliability/3276-TDD.R.](#)

Decision points in code (i.e., if statements) select the next basic block to execute; using basic blocks as a coverage metric throws away information on the length of statement sequences. The regression fit in Figure 6.45 shows a linear relationship between basic block coverage and decision coverage; the expected relationship.

Growth in coverage as the number of tests increases will depend on the distribution of input values contained in the tests and the characteristics of the code being tested.

A study by McAllister and Vouk<sup>908</sup> investigated the coverage of 20 implementations of the same specification. Two sets of random tests were generated using different selection criteria and 796 tests designed to provide full functional coverage of the specification. Figure 6.46 shows the fraction of basic blocks covered as the number of tests increases, for the 20 implementations (same color) and three sets of tests (different colors); the lines are fitted regression models of the equation:  $coverage_{BB} = a + b \times \log(tests)^c$ , where:  $a$ ,  $b$  and  $c$  are constants ( $c$  is between -0.35 and -1.7 for this example).

promised for data...?

Many software systems support multiple configurations and running the same tests on different configurations can result in different sets of statements being executed;<sup>1132</sup>

Another technique for estimating the effectiveness of a tests suite for detecting coding mistakes is to introduce known mistakes into the source and measure the percentage detected by the test suite, i.e., they cause a change of behavior that is detected by the test process. The modified source is known as a *mutant*. Ideally the characteristics of the mutations matches the characteristics of the mistakes made by the developers writing the code (mutant generation

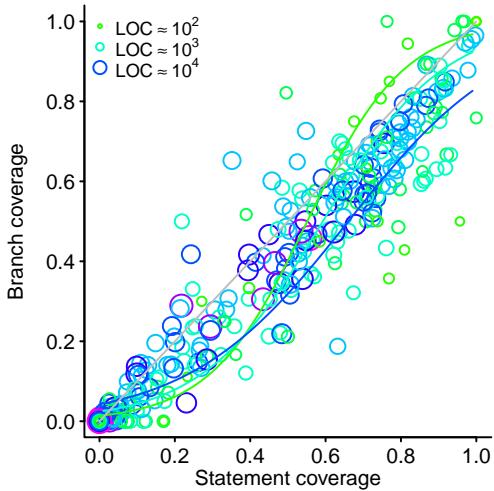


Figure 6.43: Statement coverage against branch coverage for 300 or so Java projects; colored lines are fitted regression models for three program sizes, equal value line in grey. Data from Gopinath et al.<sup>521</sup> [code](#)

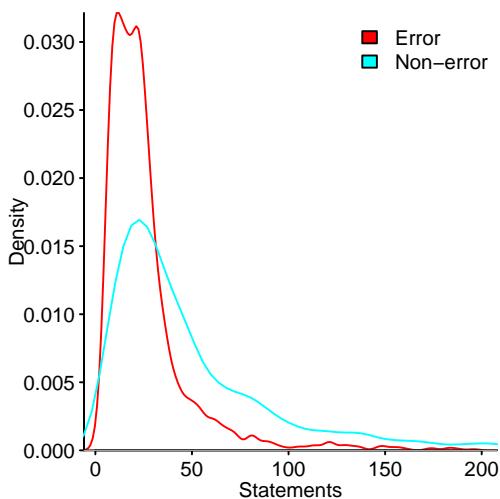
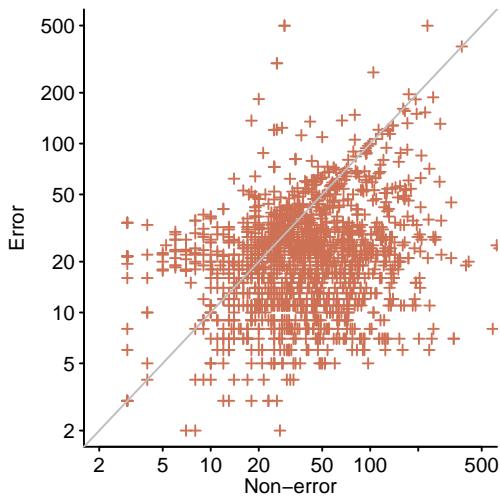


Figure 6.44: Number of statements executed along error and non-error paths within a function. Data kindly provided by Kang.<sup>734</sup> [code](#)

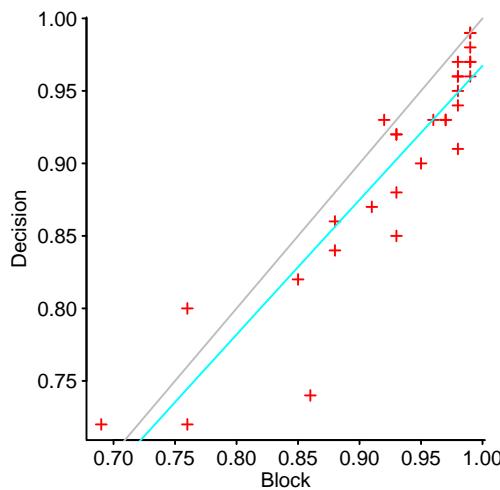


Figure 6.45: Basic block coverage against branch coverage for a 35 KLOC program. Data from Gokhale et al.<sup>508</sup> code

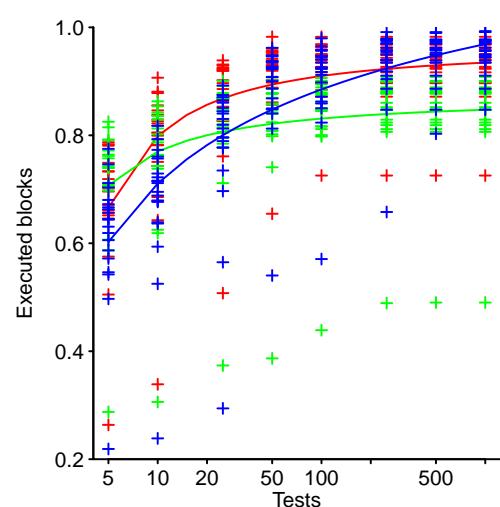


Figure 6.46: Fraction of basic blocks executed by a given number of tests, for 20 implementations using three test suites. Data from McAllister et al.<sup>908</sup> code

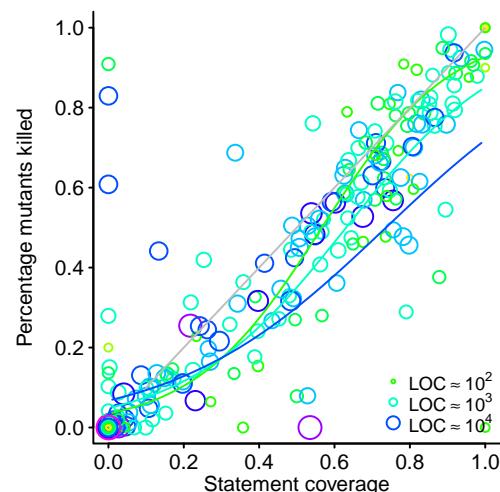


Figure 6.47: Statement coverage against mutants killed for 300 or so Java projects; colored lines are fitted regression models for three program sizes, equal value line in grey. Data from Gopinath et al.<sup>521</sup> code

techniques tend to be generic and generated mistakes have been found not to have the characteristics of developer coding mistakes<sup>522</sup>). A test suite that kills a high percentage of mutants (what self-respecting developer would ever be happy just detecting mutants?) is considered to be more effective than one that kills a lesser percentage.

Figure 6.47 shows statement coverage against percentage of mutants killed. The lines are fitted regression models, which contain a strong interaction between coverage and log KLOC... no correlation between log(loc) and test suite coverage...

A test suite's mutation score converges to a maximum value, as more mutants are tested. For programs containing fewer than 16K executable statements,  $E$ , the number of mutants needed has been found to grow no faster than  $O(E^{0.25})$ <sup>1486</sup>, a worst case confidence interval for the error in the mutation score can be calculated.<sup>520</sup>

There have been a few studies investigating the coverage of compiler validation suites<sup>257, 697</sup>...

A statement might only fail to behave as intended when the input has specific characteristics...

### 6.6.3 Cost of testing

Minimising costs by reducing the number of tests that are run...

A metric sometimes used as an indication of changes in the reliability of a software system is the number of faults discovered, per unit effort...

A study by Do, Mirarab, Tahvildari and Rothermel<sup>357</sup> investigated the cost benefit trade-offs associated with the cost of regression testing, time to market and cost of faults reported by customers...

When to stop testing software? Stopping conditions are discussed in «\_stopping\_conditions>...

It is possible to derive a relationship for the cost/benefit of continued testing vs. the cost of fixing faults experienced by customers... Stopping rules for testing after a given amount of time, based on number of faults experienced...<sup>238, 1471</sup>

### 6.6.4 Runtime issues

Techniques that have been used to mitigate against the occurrence of an error that occurs during program execution include:...

If multiple developers create implementations meeting the same specification, will coding mistakes in each program be independent of each other? Known as *N-version programming*?  
?

# Chapter 7

## Source code

### 7.1 Introduction

Source code is a form of communication, written in a programming language, whose purpose is to communicate a sequence of actions or required results<sup>i</sup> to a computer or a person.

Building a software system involves arranging source code in a way that causes the desired narrative to emerge as the program is executed with the appropriate inputs.

This chapter discusses source code from two perspectives: reducing the cognitive effort developers need to invest in processing code and how patterns of source usage can be used to validate theories of developer behavior.

People are inveterate storytellers and narrative is another way of interpreting source code...

There are many technical similarities between programming languages and human languages, however, the ways in which they are used to communicate are very different (differences have also been seen in medical imaging of brain activity<sup>448</sup>). There has been a lot of research investigating the activities involved in processing written representations of human language<sup>316</sup> (i.e., prose<sup>ii</sup>; see Section 2.5), but few scientific studies of the activities involved in processing source code. For instance, eye tracking is commonly used to study reading prose, but is only just starting to be used to study reading code (see Figure 2.38). This chapter draws on the results from studies of prose, and attempts to highlight patterns of behavior that might apply to source.

Readers may be familiar with an approach to studying language based on the products of language (e.g., words and sentences) abstracted away from their use (e.g., the work of Chomsky). The cognitive linguistics<sup>1139</sup> approach is based around how people use language, with context taken into account (e.g., intentions and social normals, such as the work of Grice<sup>550</sup>). This chapter takes a cognitive linguistics approach to source code.

- Human language use is a joint activity, based on common ground between speaker and listener<sup>iii</sup>. What a speaker says is evidence a listener uses to create an interpretation of the speaker's intended meaning. A conversation involves cooperation and coordinating activities...

Grice's maxims<sup>550</sup> provide a model of human communication, involving speaker aims and listener assumptions. Perhaps the most important aspect of human language communication is the assumption of relevance<sup>261, 1286</sup> (and even optimal relevance). A speaker says something because they believe it is worth the effort, with both speaker and listener assuming that what is said is relevant to the listener and worth the listener investing effort to understand.

---

<sup>i</sup> In a declarative language, such as SQL, the intended result is specified (e.g., return information matching the specified conditions), while code written in an imperative language specifies actions to be performed (with the author being responsible for sequencing the actions to produce an intended result).

<sup>ii</sup> English prose has been the focus of most studies, with prose written in other languages not yet so extensively studied.

<sup>iii</sup> "vaj laH: pejatlh lion ghaH yajbe' maH." Wittgenstein.

- Communicating using source code is a take it or leave it transaction, what the speaker *said* is treated as definitive by the listener, intent is not part of the process. There is no cooperative activity, and the only coordination involves the speaker modifying what has been *said* until the desired listener response is achieved.

Creating a program requires explaining everything that is relevant, including all the supporting background detail (commonly occurring narrative subplots are often supplied by library functions).

Computers do not have a conception of application assumptions and relevance. Experienced developers are aware of this and adjust their approach to communication; novices have to learn how to communicate with a listener who is not aware of their intent. One aspect of learning to program is learning to communicate with a subject who will not make any attempt to adapt to the speaker; patterns have been found in the ways novices misunderstanding the behavior of code.<sup>1067</sup>

Source code is the outcome of choices made in response to the application domain problem being solved, the algorithms used, implementation requirements, functionality provided by the programming language and the beliefs, experiences and preferences of the developers writing the code, along with a desire for short-term gratification.<sup>458</sup>

Research on application domain driven patterns of usage or high-level patterns present in programs has been thin on the ground. In the domain of compiler writing, the production-quality compiler-compiler project,<sup>837</sup> in the late 1970s, created a way of organizing optimizing compiler that has been widely used ever since.

The low-level patterns of usage found in source code arise from the many developer coding choices made in response to immediate algorithmic and implementation requirements. High-level problems need to be broken down into smaller problems, and so on down, until the developer recognises how a solution can be composed using code.

For instance, implementing the  $3n + 1$  problem requires testing whether  $n$  is odd or even. In the van der Meulen study,<sup>1389</sup> the expressions used for this test include:  $n \% 2$  and  $n \& 1$  (89% and 8% of uses respectively), along with less common sequences such as:  $(n/2)*2 == n$  and  $(n >> 1) << 1 == n$ ; around 50% of the conditions were true when  $n$  was even. Over 95% of expressions appeared as the condition of an if-statement, with most the others appearing as the first operand of a ternary operator (e.g.,  $n=(n \% 2) ? 3*n+1 : n/2$ ).

Developers have a collection of beliefs, and mental models, about the semantics of the programming languages they use, as well as a collection of techniques they have grown accustomed to using...

If, after executing the two statements: `x=1;y=x+2;`, the statement: `x=3;`, is executed, is the value of  $y$  now 5? In many programming languages the value of  $y$  remains unchanged by the second assignment to  $x$ , but in reactive programming languages<sup>86</sup> (found in spreadsheet programs) statements can express a relationship, not a one time assignment of a value (novice developers have been found to give a wide variety of interpretations of the assignment operator<sup>156</sup>).

While spreadsheets support the specification of formula and relationships between memory locations, they have not traditionally been treated as part of software engineering. One reason has been the lack of large samples of usage to study; legal proceedings against large companies are helping to change this situation.<sup>605</sup>

Acquiring an understanding of the behavior of a program, by reading its source code, is not an end in itself; one reason for making an investment to acquire this understanding, is to be able to predict its behavior sufficiently well to be able to change it. By reading source code, developers acquire beliefs about it, which are a means to an end; *understanding a program* is a continuum, not a yes/no state.

Source code is used to build programs and libraries (or packages). There are a variety of different kinds of text that might be referred to as *source code*, e.g., text that is compiled to produce an executable program, text that is used to direct the compilation process (such as Makefiles and shell scripts), text that resembles prose in configuration files...

The complexity and inter-connectedness found in software systems can also be found in other ecosystems. A study by Braha and Bar-Yam<sup>169</sup> investigated the network connections in a 16-story hospital, pharmaceutical facilities design, a General Motors vehicle design facility and

	Nodes	Edges	Average path length	Clustering coeff	Degree	Density
<b>Hospital</b>	889.00	8178.00	3.12	0.11	18.40	0.02
<b>Pharmaceutical</b>	582.00	3689.00	2.63	0.12	12.68	0.02
<b>Software</b>	466.00	1245.00	3.70	0.14	5.34	0.01
<b>Vehicles</b>	120.00	417.00	2.88	0.13	6.95	0.06

Table 7.1: Values of various attributes of the communication network graphs in various organizations. Data from Braha et al.<sup>169</sup> [code](#)

Linux. Table 7.1 lists the values of various properties of the networks and from a component network perspective software looks middle of the road.

A study by German, Adams and Hassan<sup>7</sup> investigated how developers working on the Linux kernel used Git...

### 7.1.1 Quantity of source

What is the size distribution of software systems?

A study by Landman, Serebrenik, Bouwers and Vinju<sup>808</sup> measured 13,103 Java projects. Figure 7.1 shows the number of source files, methods and sloc (within each method) for each project. A fitted regression model finds that a projects' total source lines of code is well fitted by the equation:  $sloc = e^{1.98} \text{methods}^{1.12} \text{files}^{-0.13}$

See Figure 4.21 and Figure 5.1 for information on time and costs for developing programs.

Lines of code, versions, files, support files, bytes in system shipped...

In many languages a *method* (also known as a *function*, *procedure* or *subroutine*) is the smallest self-contained unit of source code.

A study by Landman, Serebrenik, Bouwers and Vinju<sup>808</sup> measured 19K open source Java projects and the 9.6K packages (written in C) contained in a Linux distribution. Figure 7.2 shows the number Java methods and C functions containing a given number of source lines. A power law provides a good fit to both sets of data; see [sourcecode/Landman\\_m\\_ccsloc.R](#).

If we assume the projects implemented the same range of applications to equivalent levels of functionality, then differences in the distribution of method/function sizes is the result of developer use of language features, e.g., constructs that are supported and common coding idioms.

A study by Herraiz measured the number of files and SLOC in a snapshot of 3,781 open source projects. Figure 7.3 shows number of files against LOC, as a density plot;  $LOC = 161 * File^{0.99}$  is a fitted regression model. Another study<sup>607</sup> of 13 Open source projects found that the number of files was a good predictor of the number of lines of code (the multiplying factor varied from 100 to 565)...

Sometimes the only information available about a program is contained in its executable form; there have been studies<sup>1327</sup> that have built models estimating the length of the original source from the compiled machine code (the language in which code is written can have a large impact on machine code characteristics<sup>209</sup>).

The relationship between static and dynamic counts of machine code, compiled from the same source code, is affected by the processor characteristics and compiler used; Figure 7.4 shows static and dynamic percentage of call instruction opcodes, in the compiled form of various C programs (targeting various processors).

The quantity of source has been found to be an approximate predictor of compile time (e.g., one study<sup>948</sup> of Ada compilers found compile time was proportional to the square root of the number of tokens; see [sourcecode/ADA177652.R](#)). A study by Auler and Borin<sup>64</sup> investigated the time taken by a JIT compiler to generate code for functions from the SPEC CPU2006 benchmark. Figure 7.5 shows the time taken to generate machine code for 71,200 functions, containing a given number of LLVM instructions (an intermediate code). The two trends in the data are: compile time not depending on function size and compile time increasing linearly (once functions contain more than 100 instructions).

Many software systems allow some features to be optionally selected at build time. One technique for selecting which features to include in a program is conditional compilation,

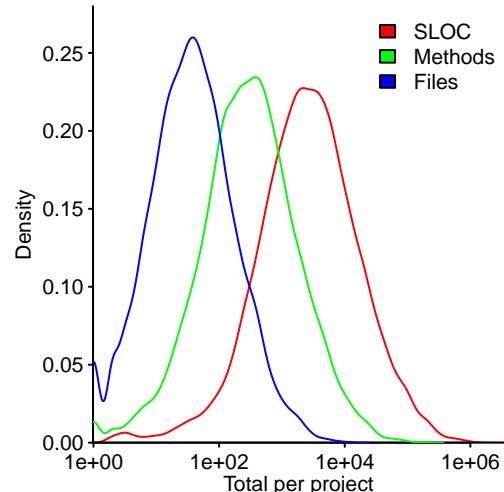


Figure 7.1: Number of source files, methods and lines of code, within methods, contained in each of 13,103 Java projects; lines are kernel density plots. Data kindly provided by Landman.<sup>808</sup> [code](#)

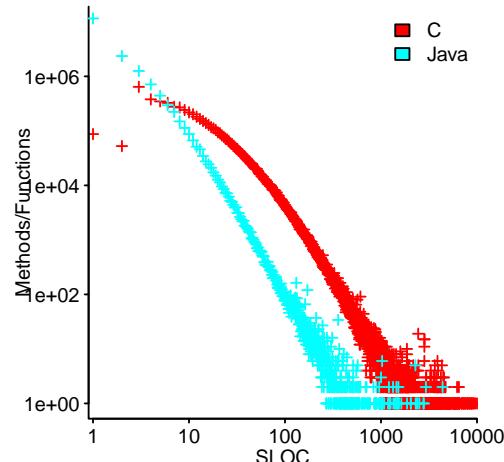


Figure 7.2: Number of methods/functions containing a given number of source lines; 17.6M methods, 6.3M functions. Data kindly provided by Landman.<sup>808</sup> [code](#)

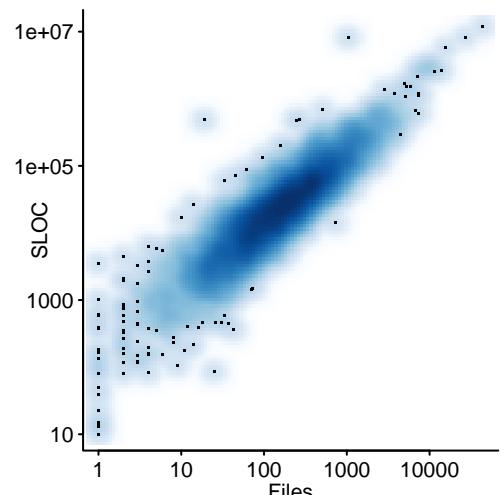


Figure 7.3: Number of files and lines of code in 3,781 projects hosted on Sourceforge. Data from Herraiz.<sup>1333</sup> [code](#)

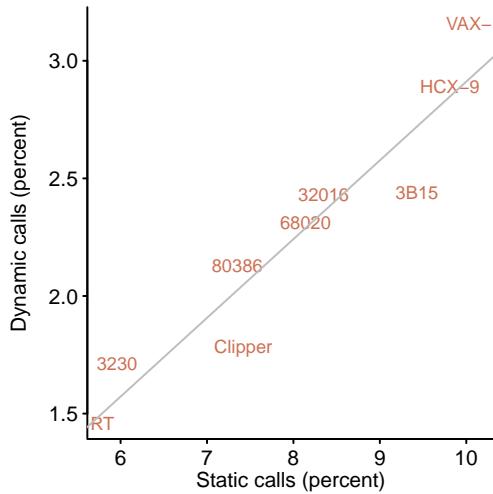


Figure 7.4: Percentage of call instructions contained in code generated from the same C source, against call execution percentage for various processors; grey line is fitted regression model. Data from Davidson et al.<sup>315</sup> [code](#)

e.g., `#if/#endif` in C and C++. A study by Liebig, Apel, Lengauer, Kästner and Schulze<sup>846</sup> measured various attributes associated with the use of conditional compilation directives in 40 programs written in C (header file contents were ignored). Figure 7.6 shows support for optional features increases with lines of code.

## 7.2 Desirable characteristics

The possibility that, at some point in the future, new code is added to existing code, or existing code is modified, provides an investment opportunity to reduce future costs by using code that has the appropriate characteristics.

- a regular turnover of staff (see Section 4.5.1) creates an incentive to reduce upfront investment in source code familiarisation costs, e.g., using constructs whose behavior is the same across implementations means that recruitment is not restricted to developers with experience of a specific implementation (this aim has to be weighed against the benefits provided by implementation specific features),
- developers' primary interaction with source code is reading it to obtain information and integrating new functionality into existing code. Increasing the memorability of previously read code reduces the time needed to reread it and reduces the cognitive effort needed to understand the behavior...
- efficiency... inefficient SQL queries<sup>175</sup>...

Questions about what code is likely to change is usually a question about what functionality is likely to change, or be added.

The terms such as *Maintainability*, *readability* and *testability* are often cited as being desirable characteristics. However, their definition is often shaped to fit the research idea or product functionality being promoted, and appear with easier, cheaper, quicker; these phrases are essentially meaningless marketing hooks.

Books, reports and memos containing recommendations about how code should be written and language constructs to avoid, so called *coding guidelines*, are something of a cottage industry, e.g., for C.<sup>306,457,468,582,641,679,744,777,912,957,1096,1097,1133,1137,1143,1220,1287,1291,1320</sup>

Stylistically, these documents are often more akin to literary criticism than engineering principles, i.e., they express personal opinions that are not derived from evidence (other than perhaps personal experience). Recommendations against the use of particular language constructs may be based on the construct being repeated involved in reported faults; however, the fault proneness of alternative constructs, that might be used are rarely included in the analysis, i.e., the current usage may be the least bad of the available options. C preprocessor,<sup>930</sup>...

Developers use language constructs because they are perceived to provide a worthwhile benefit. While it might be possible to choose between several language constructs to implement some functionality, at the time of writing there is no available evidence showing that any construct is more likely to have some desirable property, compared to another (e.g., less error prone, easier to modify or to understand by future readers of the code).

### 7.2.1 Programming languages

Do some programming languages require more or less effort from developers, to write code that has desirable characteristics?

Every programming language has its fans, people who ascribe various positive qualities to programs written in their favorite language, or in languages supporting a particular characteristic (e.g., strong typing). There is little experimental evidence for language characteristics having an impact on developer performance, and even less evidence for specific language features having a performance impact.

Figure 7.6: Number of *feature constants* against LOC for 40 C programs, with fitted regression line. Data from Liebig et al.<sup>846</sup> [code](#)

Factors that might generate a measurable difference in developer performance, when using different programming languages, include: individual knowledge and skill using the language, familiarity with the application domain, and interaction between the programming

language and problem, e.g., it may be easier to write a program to solve a problem using language X than using language Y.

Learning... Having subjects solve the same problem multiple times creates a learning opportunity, allowing subjects to produce an improved performance on each implementation. The experimental procedures employed generates information that can be used during the analysis of the data to factor this ordering effect into the created model.

Studies that compare languages using small programs<sup>993, 1389, 1501</sup> suffer from the possibility of a strong interaction between the problem being solved and the available language constructs (an effect that is likely to diminish as the size of the system grows). A study using small programs will fail to detect differences cause by differing support for programming in the large...

TODO...?

A realistic comparison of two languages requires information from many implementations of large systems targeting the same application domain.

A study by Waligora, Bailey and Stark<sup>1420</sup> compared various lines of code measurements (e.g., declarations, executable statements and code reuse) of 11 Ada and 19 Fortran NASA ground station software systems. While there were differences in patterns of behavior for various line counts, these differences could have been caused by factors not discussed in the report (e.g., the extent to which individual companies were involved in multiple projects and so in a good position to evaluate the potential for reuse of code from previous projects, or the extent to which project requirements specified that code should be written in a way likely to make it easier to reuse; see [projects/nasa-ada-fortran.R](#)).

Difference in performance between subjects and learning effects can dominate studies based on either small programs or experiments run over short intervals.

A study by Prechelt and Tichy<sup>1117</sup> investigated the impact of parameter checking of function calls (when the experiment was performed, C compilers that did not perform any checking on the arguments passed to functions, so called K&R style, were still in common use). All subjects wrote two programs, for one program using a compiler that performed the function call argument checking and for the other used a compiler that did not perform this checking. Subjects were randomly assigned to the problem to solve first and the compiler to use for each problem. The time to complete a correct program was measured.

Fitting a regression model shows that the greatest variation in performance occurred between subjects (standard deviation of 74 minutes), the next largest effect was problem ordering (with the second problem being solved 38 minutes, on average, faster than the first). The performance improvement attributed to argument checking is 12 minutes, compared to no argument checking (see [experiment/tcheck98.R](#)).

Proponents of strongly typed programming languages claim that using such languages provides benefits (e.g., fewer mistakes appear in the delivered software), compared to using languages that are not strongly typed. The following are studies for which data is available:

- Gannon.<sup>475</sup> Implemented compilers for two simple languages (think BCPL and BCPL+a string type and simple structures; by today's standards both languages were weakly typed, but one less so than the other). One problem was solved by subjects, this was designed to require the use of features available in both languages, e.g., a string oriented problem (final programs were between 50-300 lines). The result data included number of errors during development and number of runs needed to create a working program (this happened in 1977, before the era of personal computers, when batch processing was king; see [experiment/Gan77.R](#)).

There was a small language difference in number of errors/batch submissions; the difference was about half the size of that due to the experimental order of language use by subjects, both of which were small in comparison to the variation due to subject performance differences. While the language effect was small, it exists. To what extent can the difference be said to be due to stronger typing rather than only one language having built in support for a string type?

- Mayer, Kleinschmager & Hanenberg.<sup>770, 905</sup> Two experiments using different languages (Java and Groovy) and multiple problems; the performance metric was time for each subject to complete the task. There was no significant difference due to just language, but large

differences due to language/problem interaction with some problems solved more quickly in Java and other more quickly in Groovy and learning took place (i.e., the second task was completed in less time than the first). As always, large variations between subjects (see [experiment/mayerA1-oopsla2012.R](#) and [experiment/kleinschmagerA1.R](#)).

- Hoppe and Hanenberg:<sup>639</sup> Used one language (Java) and multiple problems involving making use of either Java's generic types or non-generic types. Again, the only significant language difference effects occurred through interaction with other variables in the experiment (e.g., the problem or the language ordering) and there were large variations in subject performance.

To summarise: when a language typing/feature effect has occurred its contribution to overall developer performance has been small. Possible reasons for small or non-existent, include:<sup>iv</sup> the use of subjects with little programming experience (i.e., students; experienced developers are more likely to make full use of the consistency checking provided by a type system), and small size of the programs (type checking comes into its own when used to organize and control large amounts of code).

Studies<sup>952</sup> investigating the use of a human language, to specify solutions to problems, have found that subjects make extensive use of the contextual referencing that is common in human communication. This usage, and other issues, make it extremely difficult to automatically process the specified instructions.

### 7.2.2 Understanding code

Understanding (the term *comprehension* is used in prose related research and both are used in software engineering research papers), in the context of source code, involves finding a good-enough meaning of source code to complete a given task. The process can involve at one or more levels of abstraction, e.g., a sequence of lines of code, methods in a file, classes in a program... Developers do not understand programs, they learn about them; a continuous process involving the creation of new beliefs and the modification of existing ones, with no well-defined ending.

The cost-effectiveness of understanding the complete program is likely to decrease as the size of the program increases; a study<sup>858</sup> based around a 250-line Fortran program found some developers using an as-needed strategy and others attempting to understand the complete program.

One study<sup>1074</sup> found that developers with the best performance answering questions about the code they had studied, had built an application domain model, from the code (building a domain model would require some familiarity with the application domain addressed by the 200-line program used in the study).

The beliefs acquired are influenced by existing beliefs about the programming language it is written in, general computing algorithms, and the application domain.<sup>1229</sup>

Developers sometimes refer to the *intended meaning* of source code, i.e., the meaning that the original author of the code intended to implement. Code comprehension being an exercise in obtaining an intended meaning that is assumed to exist. However, the only warranted assumption that can be made about source code is that the operations specified in it contribute to a meaning.

What can be done to reduce the effort needed to find a good-enough interpretation of code and to increase the likelihood that this information can be accurately recalled later?

Activities that appear to be very complicated, can have a simple description. For instance, some hunting rituals intended to select the best hunting location are randomization algorithms,<sup>1230</sup> whose social impact is to reduce the likelihood of over-hunting any location.

A study by Bransford and Johnson<sup>174</sup> investigated the impact of a top-level description on the amount of information subjects' remembered about a task. Try to give a meaning to the task described in the outer margin, while remembering what is involved (taken from the study):

Table 7.2 shows that subjects' recalled over twice as much information, if they were given a meaningful phrase (the topic), before reading the passage. The topic of the passage above is

	No Topic Given	Topic Given After	Topic Given Before	Maximum Score
Comprehension	2.29 (0.22)	2.12 (0.26)	4.50 (0.49)	7
Recall	2.82 (0.60)	2.65 (0.53)	5.83 (0.49)	18

Table 7.2: Mean comprehension rating and mean number of ideas recalled from passage (standard deviation in parentheses). Adapted from Bransford and Johnson.<sup>174</sup>

### Memory for language

The complexity of human languages, which have to be processed in real-time while listening to the speaker, is constrained by the working memory capacity of those involved in the communication activity.<sup>491, 591</sup> The capacity limits that make it difficult for speakers to construct complicated sentences, in real-time, create a benefit for listeners.

A study by Futrell, Mahowald and Gibson<sup>469</sup> investigated dependency length (the distance between words, in a sentence, that depend on each other; see Figure 7.7) in the use of 37 human languages. The results suggest that speakers attempt to minimise dependency length.

Sentence complexity<sup>248</sup> has a variety of effects on human performance. A study by Kintsch and Keenan<sup>757</sup> asked subjects to read single sentences, each containing the same number of words, but varying in the number of propositions they contained (see Figure 7.8). The time taken to read each sentence and recall it (immediately after reading it) was measured.

The results found that reading rate decreases, as the number of propositions in a sentence increases (with the total number of words remaining the same). Figure 7.9 shows reading time (in seconds) for: 1) sentences containing a given numbers of propositions, and 2) when a given number of propositions were recalled by subjects; with fitted regression models. Some propositions were not recalled, making the model based on recall a more accurate fit to the number of propositions that were processed. A later study<sup>758</sup> found that reader performance was also affected by the number of word concepts in the sentence and the grammatical form of the propositions (subordinate or superordinate).

### 7.2.3 Memory for material read

What do people remember about the material they have recently read?

Again, the only detailed experimental data available is from studies based on human language prose.

Studies<sup>173</sup> have found that in the short term syntax (i.e., words) is remembered, while over the longer term mostly semantics (i.e., the meaning) is remembered; explicit verbatim memory for text does occur.<sup>567</sup> Readers might like to try the following test (based on Jenkins<sup>685</sup>); part 1: A line at a time, 1) read the sentence on the left, 2) look away and count to five, 3) answer the question on the right, and 4) repeat process for the next line.

The girl broke the window on the porch.  
The hill was steep.  
The cat, running from the barking dog, jumped on the table.  
The tree was tall.  
The old car climbed the hill.  
The cat running from the dog jumped on the table.  
The girl who lives next door broke the window on the porch.  
The car pulled the trailer.  
The scared cat was running from the barking dog.  
The girl lives next door.  
The tree shaded the man who was smoking his pipe.  
The scared cat jumped on the table.  
The girl who lives next door broke the large window.  
The man was smoking his pipe.  
The old car climbed the steep hill.  
The large window was on the porch.  
The tall tree was in the front yard.  
The car pulling the trailer climbed the steep hill.

Broke what?  
What was?  
From what?  
Was what?  
What did?  
Where?  
Lives where?  
Did what?  
What was?  
Who does?  
What did?  
Broke what?  
Who was?  
The what?  
Where?  
What was?  
Did what?

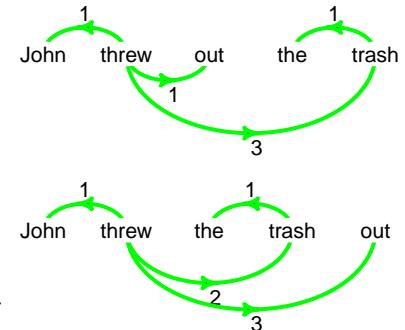


Figure 7.7: Two sentences, with their dependency representations; upper sentence has total dependency length six, while in the lower sentence it is seven. Based on Futrell et al.<sup>469</sup> code

Romulus, the legendary founder of Rome, took the women of the Sabine by force.



Cleopatra's downfall lay in her foolish trust in the fickle political figures of the Roman world.

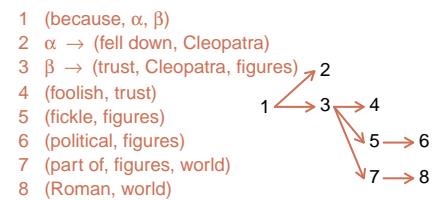


Figure 7.8: One sentence containing four and the other eight propositions, along with their propositional analyses. Based on Kintsch et al.<sup>757</sup> code

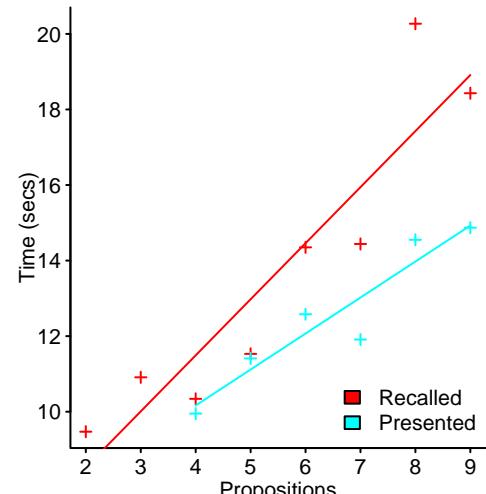


Figure 7.9: Mean reading time (in seconds) for sentences containing a given number of propositions and as a function of the number of propositions recalled by subjects; with fitted regression models. Data extracted from Kintsch et al.<sup>757</sup> code

<sup>iv</sup> Your author declares his belief that when integrated into the design process, strong typing has cost/benefit advantages.

The cat jumped on the table.  
 The tall tree in the front yard shaded the man.  
 The car pulling the trailer climbed the hill.  
 The dog was barking.  
 The window was large.

Where?  
 Did what?  
 Which car?  
 Was what?  
 What was?

You have now completed part 1. Please do something else for a minute, or so, before moving on to part 2 (which follows immediately below).

Part 2: when performing this part, do not look at the sentences above, from part 1. Now, a line at a time, 1) read the sentence on the left, 2) if you think that sentence appeared as a sentence in part 1 express your confidence level by writing a number between one and five (one expressing very little confidence and five expressing a lot of confidence in the decision) next to **old**, otherwise write a number representing your confidence level next to **new**, and 3) repeat process for the next line.

The car climbed the hill.	old____, new ____
The girl who lives next door broke the window.	old____, new ____
The old man who was smoking his pipe climbed the steep hill.	old____, new ____
The tree was in the front yard.	old____, new ____
The window was on the porch.	old____, new ____
The barking dog jumped on the old car in the front yard.	old____, new ____
The cat was running from the dog.	old____, new ____
The old car pulled the trailer.	old____, new ____
The tall tree in the front yard shaded the old car.	old____, new ____
The scared cat was running from the dog.	old____, new ____
The old car, pulling the trailer, climbed the hill.	old____, new ____
The girl who lives next door broke the large window on the porch.	old____, new ____
The tall tree shaded the man.	old____, new ____
The cat was running from the barking dog.	old____, new ____
The cat was old.	old____, new ____
The girl broke the large window.	old____, new ____
The car climbed the steep hill.	old____, new ____
The man who lives next door broke the window.	old____, new ____
The cat was scared.	old____, new ____

You have now completed part 2. Count the number of sentences you judged to be **old**.

The attribute is that all the sentences are new.

What is thought to happen is that while reading, people abstract and remember the general ideas contained in sentences (in this case, they are based on the four *idea sets*): 1) “The scared cat running from the barking dog jumped on the table.”, 2) “The old car pulling the trailer climbed the steep hill.”, 3) “The tall tree in the front yard shaded the man who was smoking his pipe.”, and 4) “The girl who lives next door broke the large window on the porch.”).

A study by Bransford and Franks<sup>173</sup> investigated subjects' confidence of having previously seen a sentence. Sentences contained either one idea unit (e.g., “The cat was scared.”), two idea units (e.g., “The scared cat jumped on the table.”), three idea units (e.g., “The scared cat was running from the dog.”), and four idea units (e.g., “The scared cat running from the barking dog jumped on the table.”). Subjects saw 24 sentences, after a 4-5 minute break they were shown 28 sentences (24 of which were new sentences), and asked to rank their confidence of having previously seen the sentence (on a 1 to 5 scale).

Figure 7.10 shows that subject's confidence of having previously seen a sentence increases with the number of idea units it contains. New sentences contained one or more idea units that were contained in previously seen sentences. The results are consistent with subject confidence level being driven by the number of previously seen idea units in a sentence, rather than the presence of new idea units.

People use their experience with the form and structure of often repeated sequences of actions, to organize the longer-term memories they form about them. The following studies illustrate the effect that a person's knowledge of the world can have on their memory for what they have read, particularly with the passage of time, and their performance in interpreting sequences of related facts they are presented with:

- A study by Bower, Black, and Turner<sup>163</sup> gave subjects a number of short stories describing various activities (i.e., scripts), such as visiting the dentist, attending a class lecture, going

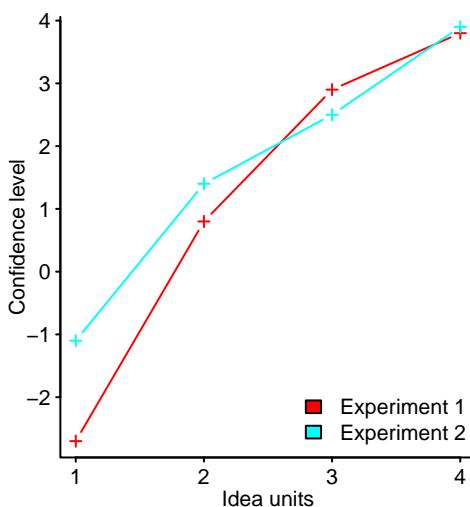


Figure 7.10: Subject confidence level, on a one to five scale (yes positive, no negative), of having previously seen a sentence containing a given number of idea units. Data extracted from Bransford et al.<sup>173</sup> code

to a birthday party, etc., to read. Each story contained about 20 actions, such as looking at a dental poster, having teeth X-rayed, etc. After a 20-minute interval, subjects were asked to recall actions contained in the stories.

The results found that around a quarter of recalled actions might be part of the script, but were not included in the written story. Approximately seven percent of recalled actions were not in the story and would not be thought to belong to the script.

A second experiment involved subjects reading a list of actions which, in the real world, would either be expected to occur in a known order or not be expected to have any order (e.g., the order of the floats in a parade). The results showed that, within ordered scripts, actions that occurred at their expected location were recalled 50% of the time, while actions occurring at unexpected locations were recalled 18% of the time at that location. The recall rate for unordered scripts (i.e., the controls) was 30%.

- A study by Graesser, Woll, Kowalski, and Smith<sup>536</sup> read subjects stories representing scripted activities (e.g., eating at a restaurant). The stories contained actions that varied in the degree to which they were typical of the script (e.g., Jack sat down at the table, Jack confirmed his reservation, and Jack put a pen in his pocket).

Memory Test	Typical (30 mins)	Atypical (30 mins)	Typical (1 week)	Atypical (1 week)
Recall (correct)	0.34	0.32	0.21	0.04
Recall (incorrect)	0.17	0.00	0.15	0.00
Recognition (correct)	0.79	0.79	0.80	0.60
Recognition (incorrect)	0.59	0.11	0.69	0.26

Table 7.3: Probability of subjects recalling or recognizing typical or atypical actions present in stories read to them, at two time intervals (30 minutes and 1 week) after hearing them. Based on Graesser et al.<sup>536</sup>

The results showed (see Table 7.3) that recall was not affected by typicality over short periods of time, but that after one week recall of atypical actions dropped significantly. Recognition performance for typical vs. atypical actions was less affected by the passage of time (e.g., subjects were asked if a particular action occurred in the story).

- A study by Dooling and Christiaansen<sup>359</sup> asked subjects to read a short biography containing 10 sentences. The only difference between the biographies read by subjects was that in some cases the name of the character was fictitious (i.e., a made up name), while in other cases it was the name of an applicable famous person. For instance, one biography described a ruthless dictator and used either the name Gerald Martin or Adolph Hitler.

After 2-days, and then after 1-week, subjects were given a list of 14 sentences (seven sentences that were included in the biography they had previously read and seven that were not included) and asked to specify which sentences they had previously read.

To measure the impact of subjects' knowledge about the famous person, on recognition performance, some subjects were given additional information. In both cases the additional information was given to the subjects who had read the biography containing the made up name (e.g., Gerald Martin). The *before* subjects were told just before reading the biography that it was actually a description of a famous person and given that persons name (e.g., Adolph Hitler). The *after* subjects were told just before performing the recognition test (they were given one minute to think about what they had been told) that the biography was actually a description of a famous person and given that persons name.

The results (see Figure 7.11) are consistent with the idea that remembering is constructive. After a week subjects memory for specific information in the passage was lost and under these conditions, recognition of sentences is guided by subjects' general knowledge. Variations in the time between reading the biography and identity of a famous character being revealed affected the extent to which subjects integrated this information.

## 7.2.4 Integrating information

Units of source code (e.g., statements and functions) are combined in a way that results in a behavioral narrative emerging during program execution. Modify an existing behavioral narrative requires that the developer acquire a good enough understanding of how execution

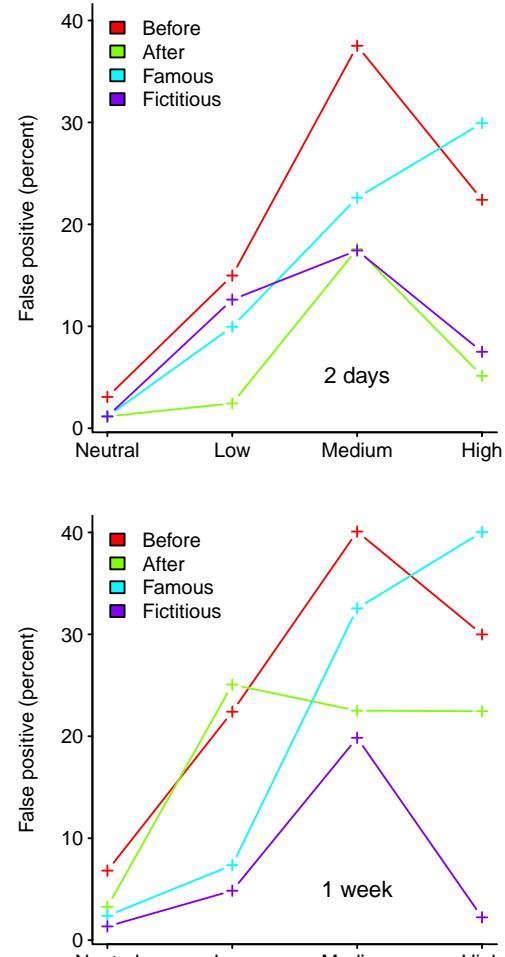


Figure 7.11: Percentage of false-positive recognition errors for biographies having varying degrees of thematic relatedness to the famous person, in *before*, *after*, *famous*, and *fictitious* groups. Data extracted from Dooling et al.<sup>359</sup> code

of the units of code combine to cause the behavior to emerge. Developers have to extract information from the units of code and integrate it into a mental model.

What are the factors that have a significant impact the cognitive effort required to integrate source code information into a mental model? Studies<sup>756,922</sup> of prose comprehension provide some clues.

The process of integrating two related items of information involves acquiring the first item, keeping it available for recall while processing other information until the second item is encountered; it is now possible to notice that the two items are related, and act on this observation.

A study by Daneman and Carpenter<sup>309</sup> investigated the connection between various measures of subjects' working memory span and their performance on a reading comprehension task. The two measures of working memory used were the *reading span* and *word span*. In the reading span test subjects have to read, out loud, sequences of sentences while remembering the last word of each sentence, which have to be recalled at the end of the sequence. The number of sentences in each sequence is increased until subjects are unable to successfully recall all the last words.

The comprehension task involved subjects reading a narrative passage containing approximately 140 words and then answering questions about facts and events described in the passage. Various passages were constructed such that the distance between the information needed to answer questions varied. For instance, the final sentence of the passage contained a pronoun (e.g., she, her, he, him, or it) that referred to a noun occurring in a previous sentence; different passages contained the referenced noun in either the second, third, fourth, fifth, sixth, or seventh sentence before the last sentence.

In the excerpt: “ . . . river clearing . . . The proceedings were delayed because the leopard had not shown up yet. There was much speculation as to the reason for this midnight alarm. Finally, he arrived, and the meeting could commence.” the question “Who finally arrived?” refers to information contained in the last and third to last sentence. The question “Where was the meeting held?” requires the recall of a fact.

The results showed a strong correlation between their performance in the reading span test and the reading comprehension test (see Figure 7.12). A similar pattern of results was obtained when the task involved listening, rather than reading. A study by Turner and Engle<sup>1372</sup> found that having subjects verify simple arithmetic identities, rather than a reading comprehension test, did not alter the results. However, altering the difficulty of the background task (e.g., using sentences that required more effort to comprehend) reduced performance.

As a coding example, given the following three assignments, would moving the assignment to `x` after the assignment to `y` reduce the cognitive effort needed to comprehend the value of the expression assigned to `z`?

```
x = ex_1 + ex_2;          /* Could be moved to after assignment to y. */
y = complicated_expression; /* No dependencies on previous statement. */
z = y + ex_1;
```

This question assumes that `ex_2` does not appear prior to the assignment to `x`, in which case there may be a greater benefit to this assignment appearing close to this usage, rather than close to the assignment to `z`.

Is reader cognitive effort reduced by having a single complex statement, rather than several simpler statements?

A study by Daneman and Carpenter<sup>310</sup> investigated subjects performance integrating information within a single sentence, compared to across two sentences, e.g., “There is a sewer near our home who makes terrific suits” (what is known as a *garden path* sentence) and “There is a sewer near our home. He makes terrific suits.” The results found that a sentence boundary can affect comprehension performance. It was proposed that this performance difference was caused by readers purging any verbatim information they held in working memory, about a sentence, on reaching its end. The availability of previously read words, in the single sentence case, making it easier to change an interpretation, based of what has already been read.

A parallel might be drawn between paragraphs in prose and function definitions in code.

A study by Kintsch, Mandel, and Kozminsky<sup>759</sup> investigated the time taken to read and summarize 1,400 word stories. The order of the paragraphs (not the sentences) in the text seen

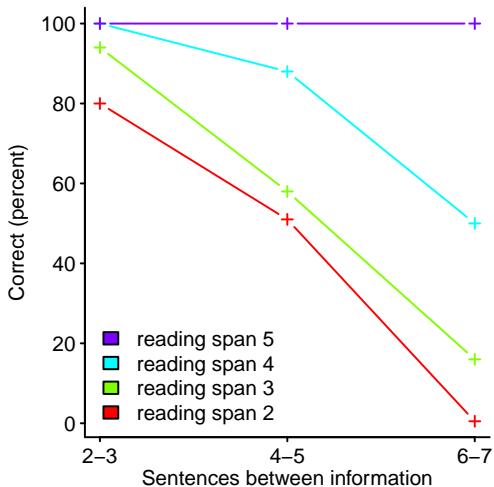


Figure 7.12: Percentage of correct responses, for subjects having a given reading span, to the pronoun reference questions as a function of the number of sentences (x-axis) between the pronoun and the referent noun. Data extracted from Daneman et al.<sup>309</sup> [code](#)

by some students was randomized. The results showed that while it was not possible to distinguish between the summaries produced by subjects reading ordered vs. randomised stories, reading time for randomly ordered paragraphs was significantly longer (9.05 minutes vs. 7.34).

slicing studies... data?...

The difference between word span and reading span as measures of working memory, is that the first is purely a measure of memory usage while the second also involves processing information (the extent to which processing information consumes, interferes with, or competes with working memory resources is a hotly debated issue). Comprehension of prose involves integrating information within and between sentences. Other studies have found that reading span correlates with performance in a number other tasks, including:

- Reading of prose involves a number of structural components, e.g., at the microstructure level words are formed from character sequences and syntactic processing of words occurs, while at the macrostructure level information has to be integrated across sentences for the narrative to be followed. A study by Graesser, Hoffman, and Clark<sup>535</sup> found that: more cognitive resources are allocated to macrostructure than microstructure processing, differences in reading speed could be attributed to different rates of microstructure processing, and variations in reader goals affect the rate at which the macrostructure is processed.
- A study by Gibson and Thomas<sup>492</sup> found that subjects were likely to perceive complex ungrammatical sentences as being grammatical. Subjects handled complex sentence that exceeded working memory capacity by forgetting parts of the syntactic structure of the sentence, to create a grammatically correct sentence.
- A study by Ehrlich and Johnson-Laird<sup>377</sup> asked subjects to draw diagrams depicting the spatial layout of everyday items specified by a sequence of sentences. The sentences varied in the extent to which an item appearing as the object (or subject, or not at all) in one sentence appeared as the subject (or object, or not at all) in the immediately following sentence. For instance, there is referential continuity in the sentence sequence “The knife is in front of the pot. The pot is on the left of the glass. The glass is behind the dish.”, but not in the sequence “The knife is in front of the pot. The glass is behind the dish. The pot is on the left of the glass.”

The results found that when the items in the sentence sequences had referential continuity 57% of the diagrams were correct, compared to 33% when there was no continuity. Most of the errors for the non-continuity sentences were items missing from the diagram drawn and subjects reported finding it difficult to remember the items as well as the relation between them.

use longer version, eRm package...

## 7.2.5 Visual organization

Source code can be visually organised in a wide variety of ways. Visual attention is a limited resource (as illustrated by Figure 7.13) and source with the appropriate layout characteristics can take advantage of the in-built hardware available to the human visual system, to reduce the conscious effort needed to visually process code, e.g., indentation takes advantage of preattentive detection of lines; Section 2.5 discussed human visual processing.

People's ability to learn means that, with practice, they can adapt to handle a wide variety of disparate visual organizations. This learning ability means that given enough practice, developers are likely to be able to learn to rapidly process many code layout rules. Using a visual organization which is already familiar to developers, reduces the start-up cost of learning to use a code layout, i.e., they can make use of prior learning to start closer to their long-term performance.

For instance, how quickly can people achieve a reading performance comparable to that achieved with a familiar layout (based on reading and error rate).

A study by Kolers and Perkins<sup>780</sup> investigated the extent to which practice improved reading performance. Subjects were asked to read pages of text written in various ways; pages contained, normal, reversed, inverted, or mirrored text, as in the following:

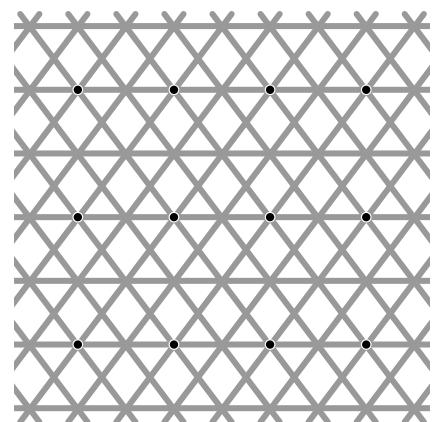


Figure 7.13: Hermann grid, with variation due to Ninio and Stevens<sup>1012</sup> to create an extinction illusion. [code](#)

- Expectations can also mislead us; the unexpected is always hard to perceive clearly. Sometimes we fail to recognize an object because we...

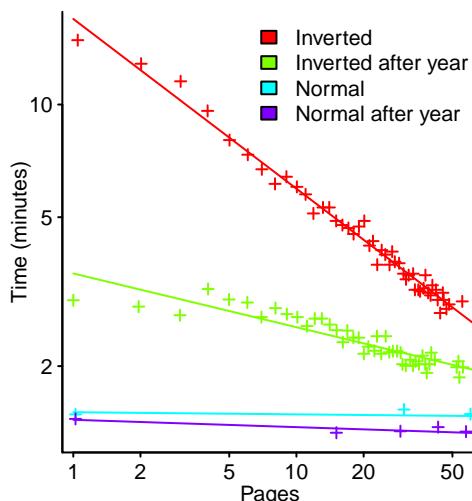


Figure 7.14: Time taken for subjects to read a page of text, printed with a particular orientation, as they read more pages (initial experiment and repeated after one year); with fitted regression lines. Results are for the same six subjects in two tests more than a year apart. Based on Kolers.<sup>779</sup> code

- \* .ekam ot detcepxe eb thgm natiruP dnalignE weN a ekatsim fo dnik eht saw tI .eb ot serad eh sa yzal sa si nam yreve taht dias ecno nosremE
- mental processes. Many other reasons can be  
These are just a few of the reasons for believing that a person cannot be conscious of all the information available.
- Several years ago a professor who teaches psychology at a large university said that a person's intelligence is limited by their ability to process information.

The time taken for subjects to read a page of text in a particular orientation was measured; the more pages subjects read, the faster they became. In a study<sup>779</sup> a year later, Kolers measured the performance of the same subjects, as they read more pages. Performance improved with practice, but this time the subjects had experience and their performance started out better and improved more quickly; these results are similar to those obtained in the letter-detection task. Figure 7.14 shows the time taken to read a page containing text in a particular orientation.

## 7.2.6 Identifier names

Identifier names provide a channel for communicating information to readers, that may reduce the effort they need to invest in understanding the source. The communications channel operates through the semantic associations triggered in the mind of the reader of the source. Readers of the source might be developers (including the original author) or tools attempting to gather and use semantic information.

The semantic associations triggered by an identifier may be traced back to information contained in the source (e.g., the declared type of an identifier), or preexisting cultural information present in developers' memory (e.g., meanings associated with words in their native language).

Human languages change over time,<sup>800</sup> with different changes sometimes occurring in different geographic communities. The names used by a community to denote entities evolves over time. One study<sup>311</sup> found a two-stage lifecycle: a linguistically innovative learning phase in which users align with the language of the community, followed by a conservative phase in which users stop responding to changes in community norms...

Source code often contains many different identifiers, each competing for developer cognitive resources. Identifiers having similar spellings, pronunciations or semantic associations may generate confusion, resulting in mistakes being made; identifiers with long names may consume cognitive resources that are out of proportion to the benefits of the information they communicate. Figure 7.15 shows the number of function definitions containing a given number of occurrences of identifiers and of distinct identifiers.

Identifiers do not appear in isolation, they occur within a context in source; one study<sup>701</sup> found that identifier names has a large impact on decisions about the precedence of binary operators in expressions.

Desirable characteristics of an identifier name include: triggering the appropriate semantic associations in a readers' mind, a low probability of being mistaken for another identifier present in the associated source code and not consuming a disproportionate amount of cognitive resources.

Identifiers are the most common source token (29% of the visible tokens in the .c files,<sup>699</sup> with comma the second most common at 9.5%), and they represent approximately 40% of all non-white-space characters in the visible source (comments representing 31% of the characters in the .c files).

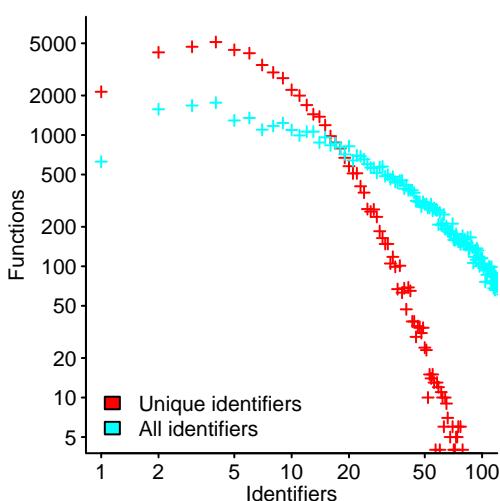


Figure 7.15: Number of C function definitions containing a given number of identifier uses (unique and all). Data from Jones.<sup>699</sup> code

# < . >	include string.h define MAX_CNUM_LEN define VALID_CNUM define INVALID_CNUM	#include <string.h> #define v1 13 #define v2 0 #define v3 1
# 13 # 0 # 1 ( * [] ) , ; * = ( ); { > ) * = ; { ( =0; < ; ++) (( []<'0' )    ( []>'9' )) * = ; } }	int chk_cnum_valid char cust_num int cnum_len cnum_status VALID_CNUM cnum_len strlen cust_num if cnum_len MAX_CNUM_LEN cnum_status INVALID_CNUM else for i i cnum_len i if cust_num i cust_num i cnum_status INVALID_CNUM	int v4(char v5[]), int *v6) { int v7, v8; *v6=v2; v8=strlen(v5); if (v8 > v1) { *v6=v3; } else { for (v7=0; v7 < v8; v7++) { if ((v5[v7] < '0')    (v5[v7] > '9')) { *v6=v3; } } }

Figure 7.16: Three versions of the source of the same program, showing identifiers, non-identifiers and in an anonymous form; illustrating how a reader's existing knowledge of words can provide a significant benefit in comprehending source code. Based on an example from Laitinen.<sup>802</sup>

Characteristics of words, in written form, where studies have found an effect on some aspect of subject performance include: *word length effect*,<sup>1002</sup> age of acquisition<sup>290,1208</sup> (when the word was first learned), frequency of occurrence<sup>69</sup> (e.g., in spoken form and various kinds of written material), articulatory features of the initial phoneme (analysis of a character sequences starts at the beginning, so...)...

Historically, most of the studies have involved English, but a growing number of large scale studies involve other languages.<sup>428</sup> Orthography (the spelling system for the written form of a language) can have an impact on reader performance, English has a deep orthography (i.e., a complex mapping between spelling and sound), while Malay has a shallow orthography (i.e., a one-to-one mapping between spelling and sound; also Spanish) and word length in Malay has been found to be a better predictor for word recognition than word frequency<sup>1473</sup> ...

Developers make use of the characteristics of their native language when creating identifier names, e.g., lexical conventions, syntax,<sup>218</sup> word ordering conventions (adjectives). Developers for whom English is a second language are likely to make mistakes, in the use of English, that are influenced by the characteristics of their native language.<sup>1330</sup>

As always, the cost of any investment a developer is willing to make to reduce the cognitive effort of subsequent readers, must be less than the total benefit obtained by all subsequent readers.

Given that most functions are only ever modified by the one developer (e.g., Figure 7.21), the primary beneficiary of any investment in local identifier names is likely to be the developer who made it.

Given that identifier communication occurs via semantic associations, what is the probability that a particular identifier will trigger the same semantic associations in multiple readers of the code?

A study by Nelson, McEvoy and Schreiber<sup>998</sup> investigated free association of words; subjects were given a word and asked to reply with the first word that came to mind. More than 6,000 subjects producing over 700,000 responses to 5,018 stimulus words.

What is the probability that the same word will be produced by more than one subject? Figure 7.17 shows the mean probability that a given percentage of subjects will produce the same word (from the same cue word). The mean percentage of unique responses was 18% (sd 9).

Table 7.4 shows the percentage of identifiers occurring in each pair of seven large software systems.

Useful semantic associations may be triggered by different words, however, the same word may trigger different semantic associations in different people?

Studies have found that people differ in the meaning they assign to the same word. For instance, the extent to which a word is thought to denote a concrete or abstract concept<sup>1084</sup> (*concrete* words, defined as things or actions in reality, experienced directly through the senses,

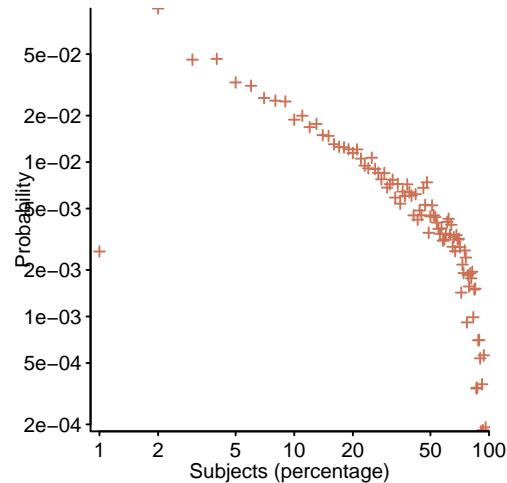


Figure 7.17: Probability (averaged over all cue words) that, for a given cue word, a given percentage of subjects will produce the same word. Data from Nelson et al.<sup>998</sup> code

whereas *abstract* words are not experienced through the senses, they are language-based with their meaning depending on other words).

	<b>gcc</b>	<b>idsoftware</b>	<b>linux</b>	<b>netscape</b>	<b>openafs</b>	<b>openMotif</b>	<b>postgresql</b>
<b>identifiers</b>	46,549	27,467	275,566	52,326	35,868	35,465	18,131
<b>gcc</b>	-	2	9	6	5	3	3
<b>idsoftware</b>	5	-	8	6	5	4	3
<b>linux</b>	1	0	-	1	1	0	0
<b>netscape</b>	5	3	8	-	5	7	3
<b>openafs</b>	6	4	12	8	-	3	5
<b>openMotif</b>	4	3	6	11	3	-	3
<b>postgresql</b>	9	5	12	11	10	6	-

Table 7.4: Percentage of identifiers in one program having the same spelling as identifiers occurring in various other programs. First row is the total number of identifiers in the program and the value used to divide the number of shared identifiers in that column. Data from Jones.<sup>699</sup>

A study<sup>1231</sup> of word choice in a fill-in-the-blank task, e.g., “in tracking the . . .”, found that probability of word occurrence (derived from large samples of language use) was a good predictor of the words chosen and the order in which subjects produced them (subjects were asked to provide 20 responses per phrase).

Readers of different natural languages will have trained on different inputs, and during school studied different subjects (each having its own technical terms). A study by Gardner, Rothkopf, Lapan, and Lafferty<sup>477</sup> asked subjects (10 engineering, 10 nursing, and 10 law students) to indicate whether a letter sequence was a word or a nonword. The words were drawn from a sample of high frequency words (more than 100 per million), medium-frequency (10–99 per million), low-frequency (less than 10 per million), and occupationally related engineering or medical words.

The results showed engineering subjects could more quickly and accurately identify the words related to engineering (but not medicine); the nursing subjects could more quickly and accurately identify the words related to medicine (but not engineering). The law students showed no response differences for either group of occupationally related words. There were no response differences on identifying nonwords. The performance of the engineering and nursing students on their respective occupational words was almost as good as their performance on the medium-frequency words.

What characteristics increase the likelihood that an identifier will be mistaken for another one?

A study by Lambert, Chang, and Gupta<sup>804</sup> investigated drug name confusion errors.<sup>v</sup> Subjects briefly saw the degraded image of a drug name. Both the frequency of occurrence of drug names and their neighborhood density were found to be significant factors in the error rate.

An analysis of the kinds of errors made, found that 234 were omission errors and 4,128 were substitution errors. In the case of the substitution errors, 63.5% were names of other drugs (e.g., **Indocin®** instead of **Indomed®**), with the remaining substitution errors being spelling-related or other non-drug responses (e.g., **Catapress** instead of **Catapres®**).

A study by Kernighan, Church, and Gale<sup>745</sup> investigated patterns of incorrect spellings; 14,742 *misspelled* words detected from 44 million words of text were used to build four tables of probabilities. For instance, of times the character *y* is deleted when it follows the character *x* (in misspelled words) divided by the number of times the character *y* follows the character *x* (in all words in the text). The other tables counted insertions, substitutions, and transpositions...

Pronunciation can be context dependent, e.g., “You can lead a horse to water, but a pencil must be lead.” and “Wind the clock when the wind blows.”

Identifiers often contain character sequences that do not match words in the native language of the reader. Studies of prose have included the use non-words, often as a performance comparison against words. Nonwords are sometimes read as a word whose spelling it closely resembles.<sup>1109</sup>

Studies of letter similarity have a long history,<sup>1356</sup> and tables of visual<sup>974</sup> (e.g., 1 (one) and 1 (ell)) and acoustic<sup>1085</sup> letter confusion have been published. When categorizing a stimulus,

<sup>v</sup> Errors involving medication kill one person every day in the U.S., and injure more than one million every year; confusion between drug names that look and sound alike account for 15% to 25% of reported medication errors

people are more likely to ignore a feature than they are to add a missing feature, e.g., **Q** is confused with **O** more often than **O** is confused with **Q**.

A study by Avidan<sup>67</sup> asked subjects to work out what a method did, timing how long it took them to reply. In the control condition subjects saw the original method and in the experimental condition the method name was replaced by xxx and local and parameter names were replaced by single letter identifiers; in all cases the method name was replaced by xxx.

Subjects took longer to reply for modified methods. When parameters were revealed first, subjects were 268 seconds slower (on average), and when locals were revealed first 342 seconds slower (the standard deviation of the between subject differences was 187 and 253 seconds, respectively; see [sourcecode/Avidan-MSc.R](#)).

The names of existing identifiers are sometimes changed.<sup>52</sup> The constraints on identifier renaming include the cost of making all the necessary changes in the code and dependencies other software may have on existing names (e.g., identifier is in a published API).

A study<sup>223</sup> of the effectiveness of two code obfuscation techniques (renaming identifiers and complicating the control flow) found that renaming identifiers had a larger impact on the time taken by subjects to comprehend and change code (see [sourcecode/AssessEffectCodeOb.R](#)).

One study<sup>628</sup> found that the time taken to find a mistake in a short snippet of code was slightly less when the identifiers were words (see [sourcecode/shorter-iden.R](#)).

## 7.2.7 Consistency

People can subconsciously learn and make use of patterns in events that occur in their environment (see Section 2.4). Consistently following a set of patterns, when writing source code, creates an opportunity for readers to make use of this ability. Individual developers do write code in a distinct style,<sup>211</sup> even if they cannot articulate every pattern of behavior they follow.

A study by Lewicki, Hill and Bizot<sup>838</sup> demonstrates the impact of implicit learning on subjects' performance, on a task that contains no overt learning component. While subjects watched a computer screen, a letter was presented in one of four possible locations; subjects had to press the button corresponding to the location of the letter as quickly as possible. The sequence of locations used followed a consistent, but complex, pattern. The results showed subjects' response times continually improving as they gained experience. The presentation was divided into 17 segments of 240 trials (a total of 4,080 letters). The pattern used to select the sequence of locations was changed after the 15th segment, but subjects were not told about the existence of any patterns of behavior. After completing the presentation subjects were interviewed to find out if they had been aware of any patterns in the presentation; they had not.

Figure 7.18 shows the mean response time for each segment. The consistent improvement, after the first segment, is interrupted by a decrease in performance after the pattern changes.

Implicit learning of grammars...

A study by Corazza, Maggio and Scanniello<sup>286</sup> investigated what they called the *coherence* between a method's implementation and any associated comment (i.e., does the method implement the intent expressed in the comment). Five Java programs (5,762 methods) were manually evaluated; coherence was positively correlated with  $\log(\text{comment\_lines})$  and negatively correlated with method LOC (see [sourcecode/SQJ\\_2015.R](#)).

A study by Buse and Weimer<sup>204</sup> investigated Computer Science students' opinions on the readability of short snippets of Java source code, rating them on a scale of 1 to 5. The students were taking first, second and third/fourth year Computer Science degree courses and postgraduates at the researchers' University.

Subjects were not given any instructions on how to rate the snippets for readability, and the outcome that individuals were trying to achieve when rating is not known, e.g., were subjects rating based on how readable they personally found the snippets to be, or rating on the basis of the answer they would expect to give if they were being tested in an exam. The study can be viewed as an investigation of implicit learning, with between student consistency of readability ratings improving as students progressed through the years of attending computer science courses.

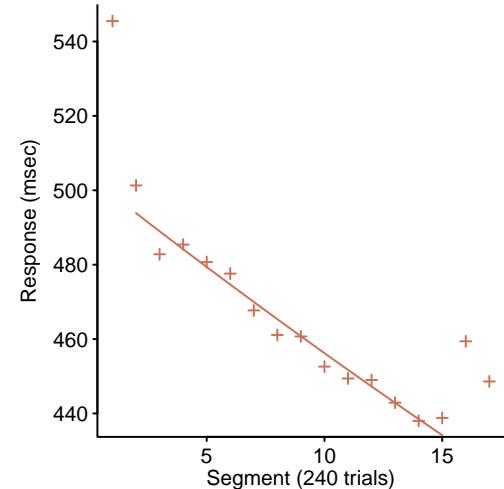


Figure 7.18: Mean response time for each of 17 segments; regression line fitted to segments 2-15. Data extracted from Lewicki et al.<sup>838</sup> code

The results show that the agreement between students readability ratings, for short snippets of code, improves as students progress through course years 1 to 4 of a computer science degree (see [developers/readability](#)).

Readability as a categorization problem...

### 7.2.8 Folklore metrics

Two source code metrics, proposed in the 1970s, have become established folklore of software development: Halstead's metric and McCabe's cyclomatic complexity metric. Use of these metrics persists, despite the studies showing them performing little better than counting lines of code. Why do developers and researchers continue to use them? Perhaps counting lines of code is considered to lack the mystique that a young field craves, or it is simply that developers are researchers have not been held accountable for inaccurate predictions.

Ptolemy's theory (i.e., the Sun and planets revolved around the Earth) was not discredited because it gave incorrect answers, but because Copernicus created a much simpler theory that produced results that eventually gave a better fit to experimental measurements (once observing equipment became more accurate).

The studies investigating what became known as Halstead's complexity metric are documented in a series of technical reports<sup>[200, 462, 523, 574, 575, 1040](#)</sup> published in the 1970s. The later reports compared theory with practice, but used very small datasets; Halstead's early work<sup>[575](#)</sup> is based on experimental data obtained for other purposes,<sup>[1501](#)</sup> and later work<sup>[462](#)</sup> used a data set containing nine measurements of four attributes (see [faults/halstead-akiyama.R](#)), others contained 11 measurements<sup>[523](#)</sup> or used measurements from nine modules.<sup>[1040](#)</sup>

Figure 7.19 shows the strong correlation between lines of code, Halstead's volume and cyclomatic complexity...

Researchers of the period found that Halstead's metric did not stand up to scrutiny: Magidin and Viso<sup>[887](#)</sup> used larger datasets (50 randomly selected algorithms...) and found many faults in the methodology in Halstead's papers; a report by Shen, Conte and Dunsmore,<sup>[1240](#)</sup> Halstead's colleagues at Purdue, written four years after Halstead's flurry of papers, contains a lot of background material and points out the lack of any theoretical foundation for some of the equations, that the analysis of the data was weak and that a more thorough analysis suggests theory and data don't agree.

An observation in the Shen et al report, that Halstead originally proposed the metrics as a way of measuring the complexity of algorithms not programs, explains the background to Halstead's uses of the term *impurities* and the discussion for the need for *purification* in his early work. Halstead points out<sup>[200](#)</sup> that the value of metrics for *algorithms written by students* are very different from those for the equivalent programs published in journals and goes on to list eight classes of impurity that need to be purified (i.e., removing or rewriting clumsy, inefficient or duplicate code) in order to obtain results that agree with the theory (i.e., cleaning data in order to obtain results that more closely agree with theory).

The paper in which McCabe<sup>[909](#)</sup> discussed what he called *complexity measures* contained a theoretical analysis of various graph-theoretic complexity measures and some wishful thinking that these might apply to software; no empirical evidence was presented. Studies<sup>[808](#)</sup> have found a strong correlation between lines of code and McCabe's complexity metric. This metric also suffers from the problem of being very easy to manipulate (see Section 6.4.3).

## 7.3 Patterns of usage

Patterns of source code usage are generated by processes such as developer beliefs and habits, application requirements, and hardware or software limits (e.g., maximum memory limits, support for a language subset).

Speakers of human languages have common patterns of word usage<sup>[828](#)</sup> and phrase usage,<sup>[137](#)</sup> the written form also has common patterns of letter usage.<sup>[706](#)</sup>

?

Individual developer patterns of usage exist,<sup>[211](#)</sup> which introduces random amounts of bias into population measurements. These *accents* come from experience with using other languages, suggestions copied from coding techniques appearing in books, course notes, etc.

Are there common patterns specific to application domains, programming language or large organizations? A study by Jones<sup>699</sup> analysed desktop computer applications written in C. While several large source code measurements projects have made their data available<sup>7</sup> and Java,<sup>544</sup> there has not been any other extensive analysis of source usage.

Patterns of usage are of interest to the extent they provide information which aids understand of the software development process. Common usage patterns are also of interest to compiler writers wanting to reduce the effort needed to implement code optimizations (i.e., by focusing on commonly occurring constructs), by static analysis tools interested in mistakes commonly made by developers, and by teachers looking to minimise what students have to know to be able to handle most situations.

Error patterns are also of interest... automate the process of locating and correcting common novice programmer mistakes<sup>694</sup>...

Runtime patterns of complete programs<sup>1157</sup> can be used to help tune the performance of both the measured program and any associated runtime system; effort might also be focused on individual language constructs.<sup>1156</sup>

A theory is of no use unless it can be used to make predictions that can be verified (or not), and any theory of developer coding behavior needs to make predictions about detectable patterns in code. For instance, given the hypothesis that developers are more likely to create a separate function for heavily nested code, then the probability of encountering an if-statement should decrease with increasing nesting depth.

Figure 7.20 shows the number of selection-statements, occurring at a given maximum nesting depth, in C code;<sup>699</sup> an if-statement is the most common form of selection-statement. The probability of encountering a selection-statement does not change with depth of nesting, implying that developers are not more likely to create a function for more deeply nested statements.

emailed, promised...

Tailored Source Code Transformations to Synthesize Computationally Diverse Program Variants<sup>104</sup>

A study by Taylor<sup>1340</sup> investigated author contribution patterns to Eclipse projects. Figure 7.21 shows the number of files modified by a given number of people; the fitted regression model is  $Files = e^{-0.87authors + 0.01authors^2}$ .

Most files are only ever edited by one person...

Pattern detection tools can add noise, or to be more exact the source code matched by tools does not always have the sought after characteristics. Also, it is not always possible to obtain definitive answers for some questions about source code characteristics, e.g., code similarity when checking for clones; tools often provide options for controlling their performance.<sup>1141</sup>

A variety of developer practices can also introduce noise into source code measurement data (apart from personal coding habits), these include the following:

- When making use of source written by third-parties, it may be more cost effective to maintain a local copy of the source files, than link to the original. A consequence of this behavior is the presence of duplicate source files in public repositories (skewing population measurements) and individual project measurements may be heavily influenced by the behavior of developers on other projects.

A study by Lopes, Maj, Martins, Saini, Yang, Zitny, Sajnani and Vitek<sup>861</sup> investigated duplicate code in 4.5 million non-fork Github hosted projects (i.e., any forks of a project were not included in the analysis). Of the 428+ millions source files written in Java, C++, Python or Javascript, 85 million were unique (percentage unique files by language...) While not explicitly labeled as forks, some projects contained a high percentage of files appearing in other projects.

Figure 7.22 shows the number of Python files containing a given number of SLOC, for a 10% sample of all 31,602,780 files and the 9,157,622 unique files.

- When modifying source, some developers commit every change they make (to the project repository), while other developers only make public commits of software they have tested (i.e., they commit changes of behavior, not changes of code).

Figure 7.23 shows the size of commits involved in fixing reported faults in Linux.

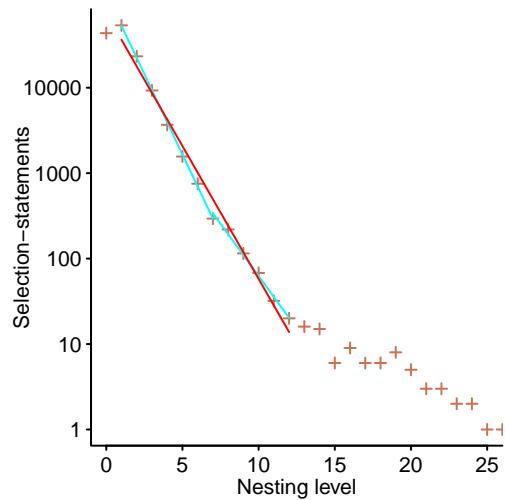


Figure 7.20: Number of selection-statements having a given maximum nesting level; for embedded C<sup>388</sup> (whose data was multiplied by a constant to allow comparison; the data for nesting depth 5 was interpolated from the adjacent points) and based on the visible form of the .c files. Data from Jones.<sup>699</sup> code

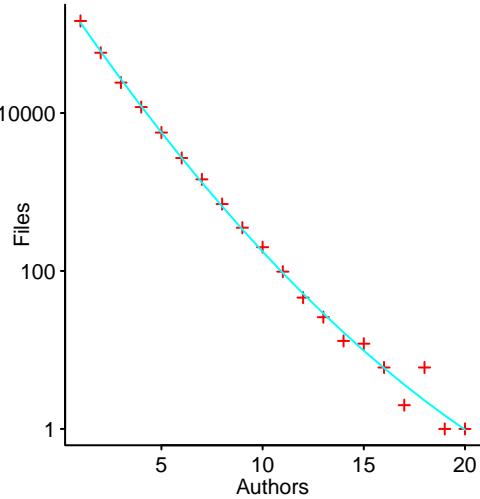


Figure 7.21: Number of files, in Eclipse projects, that have been modified by a given number of people, with fitted regression model. Data from Taylor.<sup>1340</sup> code

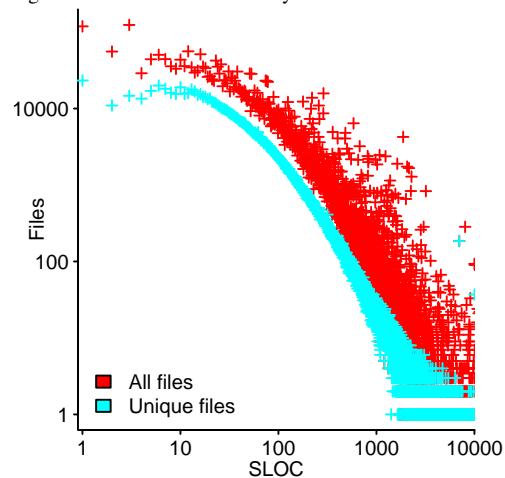


Figure 7.22: Number of Python source files containing given number of SLOC; all files, and with duplicates removed. Data from Lopes et al.<sup>861</sup> code

	Java	C++	Python	JavaScript
<b>Unique files</b>	60%	27%	31%	7%
<b>Projects</b>	1,481,468	364,155	893,197	1,755,618
<b>duplicates &gt; 50%</b>	14%	25%	18%	48%
<b>duplicates &gt; 80%</b>	9%	16%	11%	31%
<b>duplicates 100%</b>	6%	7%	6%	15%

Table 7.5: Percentage of unique files for a given language, number of projects and average percentage of duplicated files in projects in Github hosted projects written in various languages. Data from Lopes et al.<sup>861</sup>

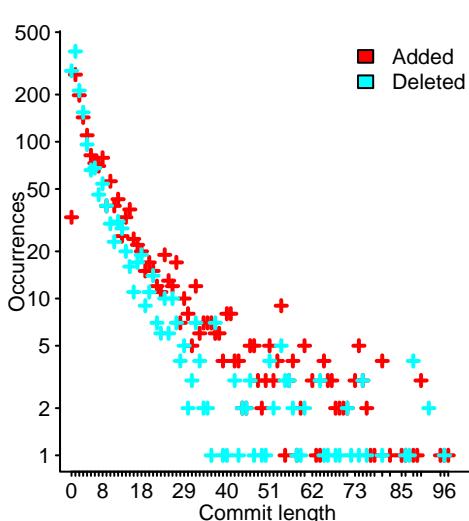


Figure 7.23: Number of commits of a given length, in lines added/deleted to fix various faults in Linux file systems. Data from Lu et al.<sup>863</sup> [code](#)

also see Figure 8.15...

A study by Kamiya<sup>728</sup> investigated how lines of code that had been deleted, were added back in to the source in a later revision... [sourcecode/iwsc2011-kamiya.R](#)

### 7.3.1 Language features

The idea that the human language we use influences our thinking is known as the *Sapir-Whorf* or *Whorfian* hypothesis.<sup>487</sup> The *strong language-based* view is that the language used influences its speakers' conceptualization process; the so-called *weak language-based* view is that linguistic influences occur in some cases, such as the following:

- *language-as-strategy*: language affects speakers performance by constraining what can be said succinctly with the set of available words; a speed/accuracy trade-off, approximating what needs to be communicated in a brief sentence rather than using a longer sentence to be more accurate,<sup>655</sup>
- *thinking-for-speaking*: for instance, English uses count nouns, which need to be modified to account for the number of items, which requires speakers to pay attention to whether one item or more than one item is being discussed; Japanese nouns make use of classifiers, e.g., shape classifiers such as hon (long thin things) and mai (flat things), and measuring classifiers such as yama (a heap of) and hako (a box of). Some languages assign a gender to object names, e.g., the Sun is feminine in German, masculine in Spanish and neuter in Russian.
- *thinking for coding* occurs when creating names for identifiers, where plurals may sometimes be used (e.g., the `rowsum` and `rowSums` functions in R),<sup>vi</sup>
- languages vary in the way they articulate numbers containing more than one digit, e.g., the components contained in 24 might be ordered as 20 + 4 (English) or 4 + 20 (French). Studies<sup>193</sup> have investigated linguistic influences on numerical cognition.

Languages vary in their complexity (i.e., there is no mechanism that ensures all languages are equally complex<sup>927</sup>).

Many programming languages in common use are still evolving, i.e., the semantics of some existing constructs are changing and support for new constructs is being added. Changing the semantics of existing constructs involves a trade-off between alienating the developers and companies currently using the language (by failing to continue to process existing code) and fully integrating new constructs into a language.

At the end of 2008 the Python Software Foundation released Python 3, a new version of the language that was not compatible with Python 2. Over time features only available in Python 3 have been back-ported to Python 2. How have Python developers responded to the availability of two similar, but incompatible languages?

A study by Malloy and Power<sup>890</sup> investigated the extent to which 50 large Python applications were compatible with various releases of the Python language. Figure 7.24 shows changes in the estimated mean compatibility of the 50 applications to 11 versions of Python, over time.

While new features are often added to languages, it is rare for a feature to be removed (at least without giving many years notice). The ISO Standards for both the Fortran and C have designated some constructs as deprecated, i.e., to be removed in the next release...

<sup>vi</sup> Your author is not aware of any studies investigating the impact of gender on the identifier names used by native speakers of different languages.

Whatever the reason for the additions, removals, changes to a language, such changes will influence the characteristics of the code written by developers. A new construct may add new functionality (e.g., atomics in C and C++), displace use of an existing construct (e.g., lambda expressions replacing anonymous classes<sup>906</sup>)...

How quickly are new languages constructs adopted and regularly used by developers? Use of language constructs depends on:

- compiler support. While vendors are quick to claim compliance with the latest standard, independent evidence is rare (e.g., compiler validation by an accredited test lab) footnote:[Your author is not aware of any studies that have investigated the delay between a feature being added to a language standard and vendor support for it; Zortech once claimed 100% C Standard compliance for their compiler (it was not illegal to claim compliance to a non-existent standard and when the standard was published their adverts switched to claiming 99% compliance a meaningless claim)].
- existing developer knowledge and practices. What incentives do existing users of a language have to learn new language features and to spend time changing existing habits? Is new language feature usage primarily driven by developers new to the language (i.e., learned about the new feature as a by-product of learning the language)?... data...

A handful of compilers now dominate the market for many widely used languages.<sup>vii</sup>

For extensive compiler driven language usage to exist, widely used diverse compilers are required (something that was once common for some languages). The small number of distinct compilers that are now widely used, means that diversity of language usage is likely to be driven by the evolution of compiler support for new language constructs and the extent to which source has been modified to adapt to changed and new features.

Studies of the introduction of generics in C#<sup>751</sup> and Java<sup>1057</sup> found that while developers make use of functionality defined using generics, in libraries, they rarely defined their own generic classes. Also, existing code in most projects is not refactored to use generics, and that the savings (in lines of code) are small.

Dates on which newly added language features were first used in a file, on a project or by a committer...

A study by Dyer, Rajan, Nguyen and Nguyen<sup>370</sup> investigated the use of newly introduced Java language features, based on the source of commits made to projects on SourceForge. Figure 7.25 shows the cumulative growth (adjusted for the growth of registered SourceForge users<sup>1477</sup>) in the number of developers who had checked in a file containing a usage of a given new feature, for the first time. Possible reasons for the almost complete halt in growth of developers using a new Java language construct for the first time include: the emptying of the pool of developers willing to learn and experiment with new language features, and developers switching to other software hosting sites (Github started in 2008).

### 7.3.2 Runtime characteristics

There is a symbiosis between cpu design and existing code; developers interested in efficiency attempt to write code that makes efficient use of cpu functionality and cpu designers attempt to optimise hardware performance for commonly occurring coding constructs (e.g., locality of reference makes it worthwhile supporting a cache). Hardware characteristics can have a significant impact on performance, e.g., Figure 13.22.

During program execution, patterns in the use of storage by distinct variables include: storage obtained via dynamic allocation usually involves relatively small objects, with program-specific sizes taking up most of the requests,<sup>895</sup> most dynamically allocated storage has a short lifetime,<sup>895</sup> and dynamically allocated objects declared using same type name tend to have similar lifetimes.<sup>1261</sup>

When execution time needs to be minimised, the relative performance of semantically equivalent coding constructs are of interest.

A study by Flater and Guthrie<sup>445</sup> compared the time taken to assign a value to an array element in C and C++, using various language constructs. The fitted regression model contained interactions between almost every characteristic measured (see [benchmark/bnds\\_chk.R](#)).

<sup>vii</sup> The availability of good enough open source compilers has led to most independent compiler vendors leaving the market. A vendor of C validation suites (selling at around \$10,000), once told your author they had over 150 licensees; a non-trivial investment for a compiler vendor.

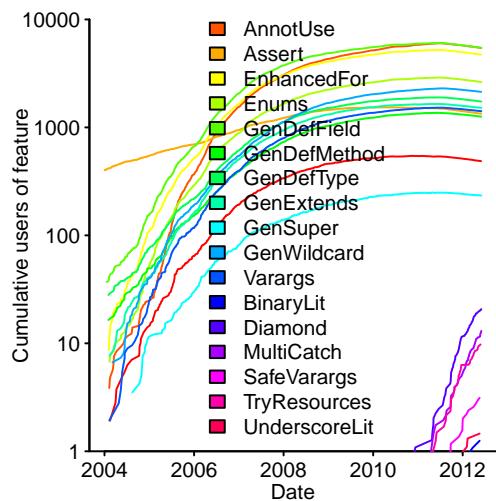


Figure 7.25: Cumulative number of developers who have committed Java source making use of new features added to the language. Data from Dyer et al.<sup>370</sup> [code](#)

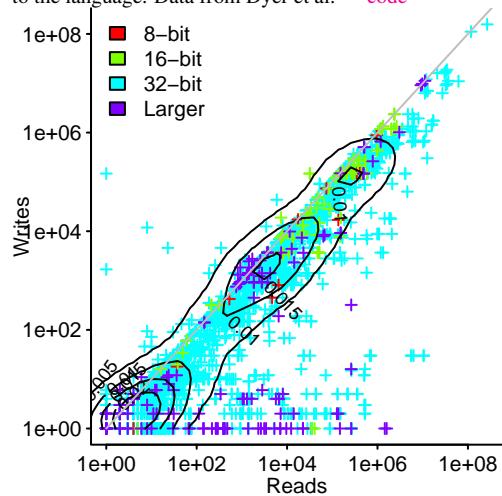


Figure 7.26: Number of reads and writes to the same variable, for 3,315 variables, made during the execution of the Mediabench suite. Data kindly provided by Caspi.<sup>220</sup> [code](#)

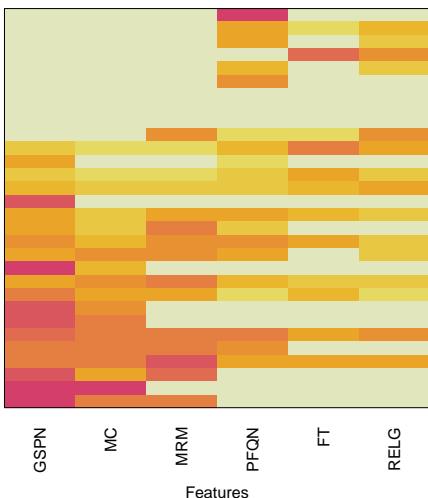


Figure 7.27: Heat map of the fraction of a file's basic blocks executed when performing a given feature of the SHARPE program. Data from Wong et al.<sup>1462</sup> [code](#)

### 7.3.3 Narrative structures

Everyday activities have a structure and include recurring themes. It is to be expected that software systems, operating within these activities, will also have a structure and include recurring themes. A study by Propp<sup>1122</sup> was perhaps the first to study recurring themes, in his analysis of Russian folktales.

A few studies have attempted to find narratives in software, e.g., so called *programming plans*.<sup>1459</sup> Computer games implemented using an Interactive Fiction approach contain source code at a high level of abstraction; the code narrative is essentially the game narrative.<sup>900</sup>

The functionality provided by a program emerges from the execution specific sequences of code, and this code may be scattered over the complete source used to build a program.

A study by Wong, Gokhale and Horgan<sup>1462</sup> investigated the execution of basic blocks, within the 30 source files making up the SHARPE program, when fed inputs that exercised six of the supported features (in turn). Figure 7.27 shows the fraction of basic blocks in each file executed when processing input exercising a particular feature.

Modularity and information hiding are two commonly used techniques for partitioning the code narratives. The costs and benefits of using this approach are...

In biological systems connections between components have a cost (e.g., they cause delays in a signalling pathway) and modularity is a technique for reducing connection costs;<sup>266</sup> modularity as a characteristic that makes it easier to adapt more quickly when the environment changes may be an additional benefit, rather than the primary driver towards modularity.

When processing source, readers have to assume that everything is relevant, until proven otherwise. Modularity provides a means of reducing the amount of information that needs to be processed by readers of the code... Prior to modifying code, the impact of the changes need to be understood. When uses of a variable are scattered about the source, ...

Localising the impact of mistakes... Reduces the likelihood that a mistake in one part of the code will impact other parts of the code...

As the amount of functionality supported by a program grows, it may become more profitable to build targeted configurations of the program (i.e., programs supporting a subset of the available features), rather than one program supporting all features. Reasons for supporting multiple configurations include targeting a particular market segment and reducing program resource usage, e.g., the likely cpu time and memory consumed when running the program with the options enabled/disabled.<sup>1252</sup>

Rather than maintaining multiple copies of the source, conditional compilation may be used to select the code that is included in a particular program build. One form of conditional compilation directive tests whether an identifier has been defined as a macro; a program feature is made optional by defining (or not) the corresponding identifier (sometimes known as a *feature test macro*, *feature constant*, or *build flag*). Figure 7.6 shows the number of unique *feature constants* appearing in programs containing a given number of lines of code. emailed, promised...

A study by Berger, She, Lotufo, Wąsowski and Czarnecki investigated the interaction between build flags and optional program features. Figure 7.28 shows the number of optional features that are enabled when a given number of build flags are set (also see Figure 11.27).

How extensive is the impact of build flags on source code? A study by Ziegler, Rothberg and Lohmann<sup>1498</sup> investigated the number of source files in the Linux kernel affected by configuration options. Figure 7.29 shows the number of files affected by the cumulative percentage of configuration options; the impact of 37.5% of options is restricted to a single file and some options have an impact over tens of thousands of files.

Figure 7.28: Number of optional features selected by a given number of flags. Data kindly provided by Berger.<sup>129</sup> [code](#)

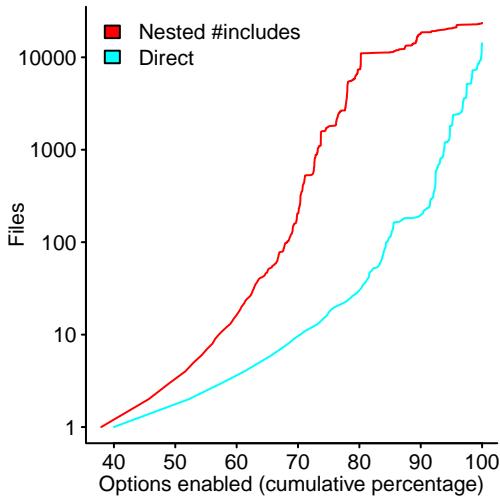


Figure 7.29: Cumulative percentage of configuration options impacting a given number of source files in the Linux kernel. Data kindly provided by Ziegler.<sup>1498</sup> [code](#)

### 7.3.4 Control flow

The ordering of operations is fundamental to producing the desired program behavior. Control flow statements specify possible orderings of code execution. acquire details about control flow and data flow<sup>1075</sup>

An if-statement represents a decision point, its presence comes from either an application requirement or an internal house-keeping issue (e.g., an algorithmic requirement or checking for an error condition).

Figure 7.30 shows...

Loop statements are where most programs spend most of their time executing (the characteristics of code within loops is intensively studied by compiler writers, optimizing loop code has the greatest cost/benefit).

The **goto** statement was the first form of control flow. The introduction of so called *structured programming*, along with the spread of language support for other forms of control flow statements, saw a significant reduction in its use. The **goto** statement in Ada and C is mostly used to jump out of nested constructs, to statements appearing either earlier or later,<sup>485, 699</sup> one study<sup>986</sup> found that 80% of usage in C was related to error handling.

Exception handling (known as *signal handling* in some languages) often involves control flow between functions/methods. Use of exception handling in C# and Java...

emailed for data...

Many languages support a **switch** statement to handle the situation where multiple conditions may be true at some point. In some cases, developers may choose to use a sequence of if-statements, e.g., the following is from the source of a version of grep:

```
if ((c = *sp++) == 0)
    goto cerror;
if (c == '<') { ... }
if (c == '>') { ... }
if (c == '[') { ... }
if (c == ']') { ... }
if (c >= '1' && c <= '9') { ... }
```

The ordering of if-statements may be driven by developer beliefs about the most efficient order to perform the tests, always adding new tests last, or some other reason. Evolution of control flow usage, a study by Jones<sup>702</sup>...

A study by Yang, Uh, and Whalley<sup>1470</sup> used profile information to reorder if-statements, by frequency of their conditional expression evaluating to true. The average performance improvement, on a variety of Unix tools, was 4%.

### 7.3.5 Statements

Statements have been the focus of the majority of studies of source code...

emailed for data...

Some languages support the creation of executable code at runtime, e.g., concatenating characters to build a sequence corresponding to an executable statement and then calling a function that interprets the string just as-if it appeared in a source file.

A study by Rodrigues and Terra<sup>1170</sup> investigated the use of dynamic features by 28 Ruby programs. On average 2.6% of the language features appearing in the source were dynamic features; it was thought possible to replace 50% of the dynamic statements with static statements (a consequence of developer habit of using template driven solutions...).

Figure 7.31 shows the number of dynamic statements, LOC, and methods appearing in Ruby programs containing a given number of those constructs. Lines are a power law fit, with the exponent varying between 0.8 and 0.9.

#### 7.3.5.1 Literal values

Literal values appear in the source of a variety of reasons, including: required by the algorithm,<sup>53</sup> size or number of elements in declarations of array types,<sup>699</sup> implementations specific values (e.g., urls, dates, etc), application domain values and personal choices of developers (for a discussion of some of the factors that influence the numeric values chosen by people, see Section 6.4.1).

What is the distribution of numeric values in application domains? Figure 7.32 shows the yearly occurrence of number words (averaged over each year since 1960) in Google's book data. The English counts are larger because most of the books processed were written in

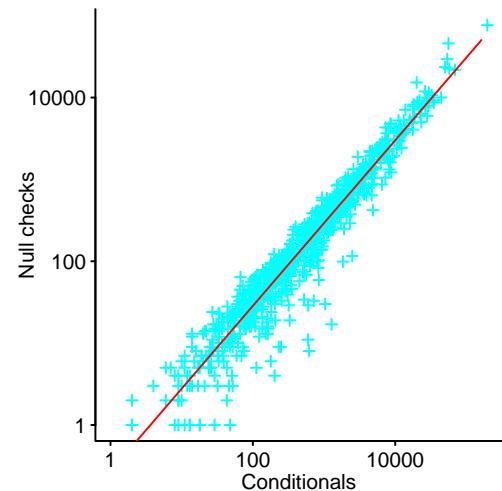


Figure 7.30: Total if-statements and if-statements whose condition involves a null check, in each of 800 Java projects. Data kindly provided by Osman.<sup>1039</sup> code

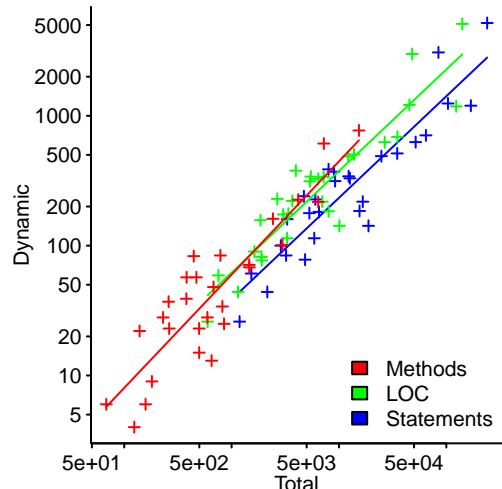


Figure 7.31: The number of dynamic statements, LOC and methods against total number of those constructs appearing in 28 Ruby programs; lines are power-law regression fits. Data from Rodrigues et al.<sup>1170</sup> code

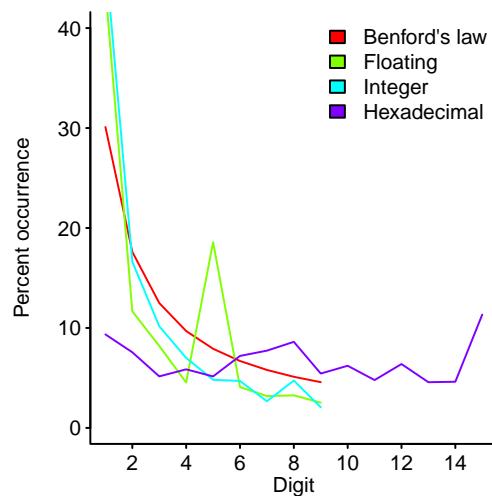


Figure 7.33: Percentage occurrence of the most significant digit of floating-point, integer and hexadecimal literals in C source code. Data from Jones.<sup>699</sup> [code](#)

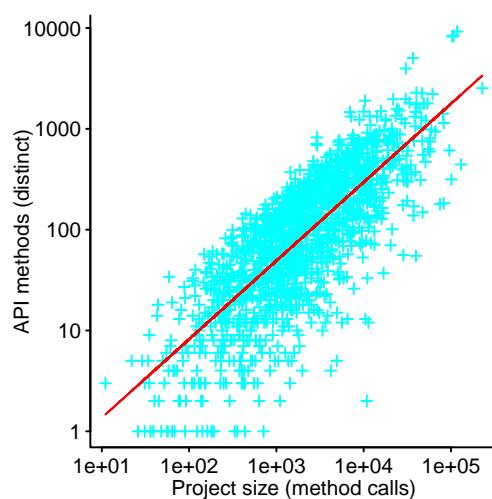


Figure 7.34: Number of distinct API methods called in 1,435 Java projects containing a given number of method calls, with regression fit. Data from Lämmel et al.<sup>805</sup> [code](#)

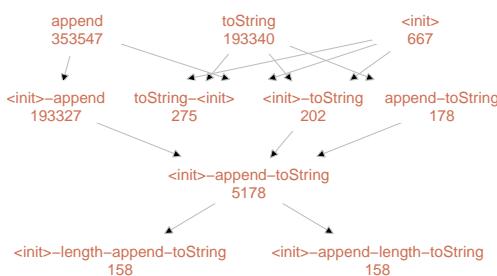
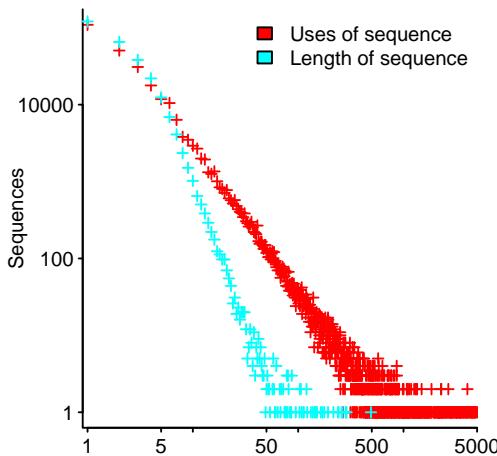


Figure 7.35: Sequences of methods, from `java.lang.StringBuilder`, called on the same object; based on 3,418 Jar files. Data from Mendez et al.<sup>933</sup> [code](#)



English. Decade values (e.g., "ten", "twenty") follow their own trend which is much higher than surrounding values.

Figure 13.3 shows that the distribution of numeric literals varies with the representation used to specify values.

Repeated occurrence of a small set of values during execution...

### 7.3.5.2 Calls

Roughly 1-in-5 statements contains an explicit function/method call (compilers sometimes introduce additional calls to implement language constructs; see Figure 7.4).

How many distinct method calls occur within a project?

A study by Lämmel, Pek and Starck<sup>805</sup> investigated calls to methods from third-party APIs and those defined within 1,435 projects. Figure 7.34 shows the number of distinct API methods called against project size (measured in method calls); the line is the fitted regression model:  $unique = e^{-1.5} calls^{0.78}$ .

The function/method called may be passed as an argument (i.e., a *callback*). A study by Gallaba, Mesbah and Beschaastnikh<sup>472</sup> investigated the use of callbacks in 130 Javascript programs. Figure 7.37 shows the total number of calls in each program, against the number of calls containing callbacks and just anonymous callbacks; lines are the fitted regression models:  $allCallbacks = e^{-0.8} allCalls^{0.86}$  and  $anonCallbacks = e^{-1.5} allCalls^{0.8}$ , respectively.

Some calls can be combined into a narrative sequence, e.g., open a file, write to it and then close it. Detecting some such narrative sequences is straightforward in object-oriented software; when a variable refers to an object, methods associated with the object can be called using the variable, e.g., `var.strLength()`.

A study by Mendez, Baudry and Monperrus<sup>933</sup> investigated method sequences called on the same variable, based on an analysis of 4,888 classes in 3,418 Jar files (i.e., Java bytecode; some method calls may not appear explicitly in the original source, e.g., the compiler maps string concatenation using binary + to a call to the method `StringBuilder.append`). Figure 7.35 shows the 10 most frequent sequences of `java.lang.StringBuilder` methods called on the same variable (lines connect methods forming a sequence; method call argument types were ignored).

Figure 7.36 shows the number of sequences having a given length (i.e., measured in methods) and number of sequences that appear in the code (i.e., are used) a given number of times...; only classes whose method sequences are used at least 100 times are included.

### 7.3.5.3 Use of variables

Patterns in the use of variables is a rarely studied subject.

Many variables take one of approximately 10 roles.<sup>1194</sup> Roles include: *stepper* which systematically takes predictable successive values, *follower* which obtains its new value from the old value of another variable, and *temporary* which holds some value for a short time.

Some usage patterns are driven by the constructs available in the language used. For instance, in C loops often involve three accesses of the loop control variable, e.g., `for (i=0;i<10; i++)`; while in languages that support constructs of the form `for (i in v_list)`, only one access is required; in languages with vector operations loops may not explicitly appear in the source.

A study<sup>396</sup> of four large PHP applications found that less than 1% of variables were assigned values having different types (e.g., assigning an array and later assigning an integer).

An analysis<sup>272</sup> of integer usage in C found that around 20% of accesses to variables having an integer type were made in a context having a signedness that is different from the declared type; also the declaration of variables having an integer type are not usually modified.

### 7.3.6 Declarations

Declarations provide two basic services: a method of specifying the representation to be used for values, e.g., a variable's type, and a method of associating one or more names with stored values.

How do declarations change over time?

A study<sup>997</sup> of the release history of a number of large C programs, over 3-4 years and a total of 43 releases, found that one or more fields were added to one or more existing structure or union types in 79% of releases, while structure or union types had one or more fields deleted in 51% of releases; another study<sup>7</sup> found one or more existing fields had their types changed in 35% of releases. [glob-var-usage.sxc](#)

A study by Robbes, Röthlisberger and Tanter<sup>1163</sup> investigated data extensions (i.e., the visitor pattern) and operation extensions to Smalltalk classes; the 2,505 projects analyzed contained 95,662 classes, forming 48,595 class hierarchies (with 41% containing more than one class). Figure 7.38 shows the number of data and operation extensions made to 1,560 class hierarchies containing both kinds of extension. The fitted regression model is:  $\log(data) = 0.91 + 0.14 \log(operation)^2$ .

#### 7.3.6.1 Visibility

Large programs contain tens of thousands of identifier uses, and various language constructs have been created to allow developers to manage visibility and accessibility of declared identifiers, e.g., the private and public attributes in Java; R is an example of a language that provides no such functionality.

There are many reasons why identifiers may have greater visibility than necessary, e.g., classes and methods in Java<sup>1407</sup> (see [sourcecode/TR\\_DCC-overExposure/TR\\_DCC-overExposure.R](#)), including: at some point in the past greater visibility was required, developers are not aware of the situation, future plans...

#### 7.3.6.2 Ordering of definitions in aggregate types

Patterns have been found in the ordering of declarations within Java class and C struct types; members sharing an attribute are often grouped together or have a preferred relative ordering.

These usage patterns may be the result of many developers making individual choices (either explicitly or implicitly), or similar ordering rules are specified (and followed), e.g., the Java coding conventions (JCC)<sup>1326</sup> specifies a recommended ordering for the declaration of: class variables (or fields), instance variables (or static initializers), constructors and methods (see Figure 12.5).

Section 12.3 discusses the analysis of items thought to have a preferred order.

Patterns in the ordering of fields in C struct types are discussed in Section 9.6.1. One interpretation of the pattern sound is reducing unused storage by placing together objects whose types have the same alignment requirements.

A study by Geffen and Maoz<sup>484</sup> investigated various patterns of method ordering within Java classes; for instance, the rules specified by the StyleCop tool (i.e., grouped by access modifiers), the commonly seen pattern of called methods appearing after the method that involved them, and the concept of connectivity (i.e., related methods appearing close to one another)...

A study by Biegel, Beck, Hornig and Diehl<sup>138</sup> investigated the impact of the kind of activity performed by a method on its relative ordering; the method activity attributes considered were: 1) all static methods (declared using the static keyword), 2) initializers (method name begins with init), 3) getters and setters (non-void return type and method name begins with get, is or set followed by a capital letter) and 4) all other non-static methods.

Figure 7.39 shows that methods performing two of the activities are very likely to be ordered before methods performing the remaining two other activities. Method declaration sequences containing more than two kinds of method activity occurred in 62% of contexts analysed, with 54% of these passing the threshold needed for analysis, leaving 33% of all declarations sequences.

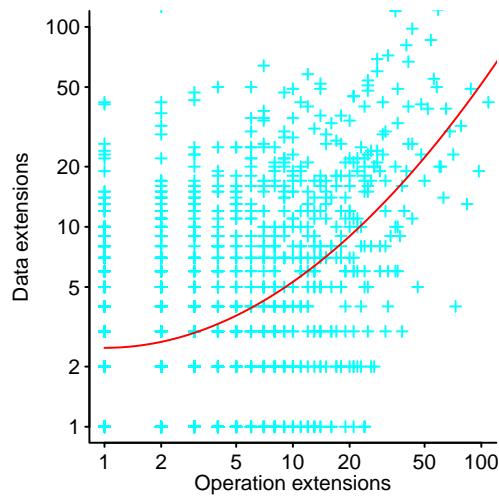


Figure 7.38: Number of data and operation extensions to 1,560 Smalltalk class hierarchies that contain both kinds of extension; line is a fitted regression model. Data from Robbes et al.<sup>1163</sup> [code](#)

The original author may not be responsible for maintenance and it is possible that declarations added during maintenance were not inserted into an existing structure/class declaration according to the ordering pattern used during initial development; experience shows that some developers prefer to add new declarations at the end of an existing structure/class.

### 7.3.6.3 Unused variables

Source sometimes contains declarations of identifiers which are not referenced outside of the declaration. Reasons for unused variables include: a mistake has been made (e.g., the variable should have been referenced), the variable was once referenced (i.e., the declaration is now redundant), and there is an expectation of future need for the variable.

Function parameters are an example where deleting the declaration of an unused variable may require a greater investment than the developer working on the function is willing to make, i.e., changing the number of function parameters requires corresponding changes to the arguments of all calls.

Figure 7.40 shows the total number of functions having a given number of parameters, a given number of unused parameters, and various fitted regression models. Unused function parameters, which at around 11% of all parameters are slightly more common than unused local variables.

The fitted regression model, for the number of functions containing a given number of unused parameters is:  $functions = e^{4.6 - 0.5 \times unused}$ . In practice, a function is likely to acquire unused parameters one at a time, as the code evolves. An alternative formula for estimating the number of functions containing  $u$  unused parameters is:  $\sum_{p=u}^8 \frac{F_p}{7p}$ , where  $F_p$  is the total number of function definitions containing  $p$  parameters.

## 7.4 Building programs

Analysis of Makefile usage<sup>899</sup>...?

size C/C++ systems, between 83% and 97% recompliations, specified in makefiles, were unnecessary<sup>1488</sup>...

The components of a system often need to be built in a specific way... emailed...

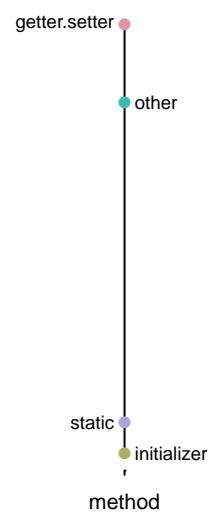


Figure 7.39: "Worth estimate" for the kind of method activity attribute (see Section 12.3). Data from Biegel et al.<sup>138</sup> [code](#)

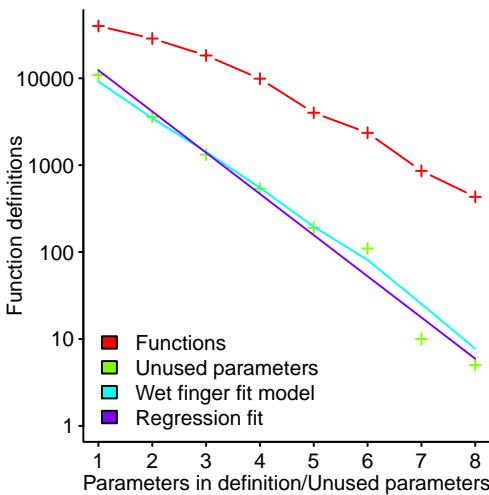


Figure 7.40: Number of C functions having a given number of unused parameters; straight lines are fitted regression models. Data from Jones.<sup>699</sup> [code](#)

# Chapter 8

## Stories told by data

### 8.1 Introduction

An appreciation of the general patterns present in the data is the starting point for understanding the stories the data has to tell and this chapter starts with a discussion of the techniques that might be used to uncover these patterns. The person performing the analysis may have to communicate the story they have discovered to other people and the second part of the chapter deals with communicating stories about data.

Ideally you will have a clear idea of the questions for which answers are sought, in practice there may be a lot of uncertainty about exactly what questions are.

Ideally you will have the time and resources needed to obtain the data needed to answer the questions asked; obtaining data is often time-consuming and/or expensive and it is often necessary to make do with whatever data is cheaply and quickly available (even if it only indirectly relate to the questions being asked).

Ideally the data contains no noise, in practice the available data may be very noisy and cleaning may be very time-consuming.

Ideally the statistical analysis techniques used is capable of providing answers to the desired level of certainty, in practice it may not be possible to draw any meaningful conclusions from the data or more questions will be uncovered.

Data analysis is like programming in that you get better with practice, there are a few basic techniques that can be used to solve many problems and doing what you did on a previous successful project can save lots of time.

This book assumes that you have a dataset and collecting data is only discussed as a secondary issue.

Perhaps the information contained in the measurements is not good and need to presented in a favourable light. This book aims to improve understanding, not obscure understanding.

There are many ways in which information can be presented in a way that misleads. This books tries to highlight techniques that aid understanding and without empirical data on the most common mistakes there is nothing to guide any discussion.

At a bare minimum, the story told by an analysis of data needs to meet the guidelines for truthfulness in advertising specified by the national advertising standards' authorities. It manufacturers of soap powder have to meet these requirements in what they tell the public, then so should you.

**Check assumptions derived from visualizations** Assumptions suggested by a visualization of a data need to be checked numerically. For instance, Figure 8.1 shows professional software development experience, in years, of subjects taking part in an experiment using a particular language; it suggests that as a group, the PHP subjects are more experienced than the Java subjects.

Comparing years of experience for the PHP and Java developers, using a permutation test, shows that the difference in mean values is not significant (there are only nine subjects in each group and the variation in experience is within the bounds of chance; see [communicating/postmortem-answers.R](#)).

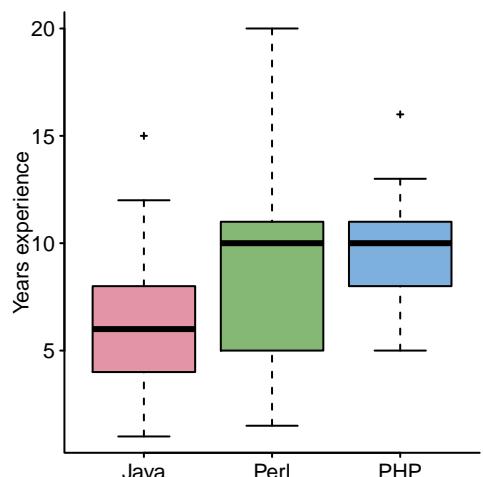


Figure 8.1: Years of professional experience in a given language for experimental subjects. Data from Prechelt.<sup>[115]</sup> code

## 8.2 Finding stories in data

You have some data. Perhaps you are expecting to see a particular pattern of behavior (e.g., an exponential relationship between two variables), perhaps you have some questions you would like answered or perhaps you would like to know what, if any, *interesting* stories the data has to reveal.

This section covers initial data exploration. Examples of more detailed analysis that might be performed after a basic appreciation has been achieved are covered in subsequent chapters. The table of contents and index can be used to locate possible techniques to analyse whether your data contains the expected pattern of behavior.

If you have questions you would like answered, then you will need to translate the questions into expected patterns of behavior that can be extracted from the data. Of course there is never any guarantee that the data contains the information needed to answer any of the questions you have.

If you don't have specific questions, you can explore the data looking for commonly occurring patterns and then try to weave anything you find into a consistent story.

Figure 8.2 shows some common and less commonly seen patterns in data. The left column shows data forming lines of various shapes; a straight line is perhaps the most commonly encountered pattern in data and points may all be close to the line or form a band of varying width. The right column shows data clustering together in various regular shapes. Uncovering a pattern is the next step along the path to understanding the processes that generated the sample measurements.

Interestingness is in the eye of the beholder. Reading through the data analysis examples in this book will give you some idea of the patterns that might be found in data and hopefully suggest some questions whose answer are of interest to you; if you still don't have a question to ask of your data and require an answer, then you might as well accept 42.

**Compelling.** Sometimes the numbers are so compelling that statistical analysis seems unnecessary. For instance, relative spacing is sometimes used within the visible form of expressions to highlight the relative precedence of binary operators (e.g., more whitespace around the addition operator when it appears adjacent to a multiplication, e.g.,  $5 + 2*3$ ). Table 8.1 shows that when relative spacing is used it nearly always occurs in a form that gives the operator with higher precedence greater proximity to its operands (relative to the operator having lower precedence). The number of cases where the reverse occurs is so small, that the either the developer who wrote the code did not know the correct relative precedence or there is a fault in the code.

	Total	High-Low	Same	Low-High
<b>no-space</b>	34,866	2,923	29,579	2,364
<b>space no-space</b>	4,132	90	393	3,649
<b>space space</b>	31,375	11,480	11,162	8,733
<b>no-space space</b>	2,659	2,136	405	118
<b>total</b>	73,032	16,629	41,539	14,864

Figure 8.2: Plots of sample values having various visual patterns. [code](#)

Table 8.1: Number of expressions containing two binary operators having the specified spacing (i.e., no spacing, no-space, or one or more whitespace characters (excluding newline), space) between a binary operator and both of its operands. High-Low are expressions where the first operator of the pair has the higher precedence, Some are expressions where the both operators of the pair have the same precedence, Low-High are expressions where the first operator of the pair has the lower precedence. For instance,  $x + y*z$  is space no-space because there are one or more space characters either side of the addition operator and no-space either side of the multiplication operator, the precedence order is Low-High. Based on the visible form of the .c files. Data from Jones.<sup>699</sup>

A study by Landy and Goldstone<sup>811</sup> found that subjects were more likely to give the correct answer (and answer more quickly) to simple arithmetic expressions when there was greater visual proximity between the operands separated by the binary operator having the higher precedence.

### 8.2.1 Initial data exploration

Initial data exploration starts with the messy subject of how data is formatted (lines containing a fixed number of delimited values is the ideal form because lots of tools can accept this as input, if a database is provided it may be worth extracting the required data into this form).

A programmer's text editor is probably the best tool for an initial look at data, unless the filename suggests it is a known binary format (such as a spreadsheet or database). For data held in spreadsheets exporting the required values to a csv file is often the simplest solution.

This initial look at the data will reveal some characteristics of the values measured, such as:

- number of measurement points (often the number of lines) and number of attributes measured (often the number of columns),
- what kind of attributes have been recorded (e.g., date, time, lines of code, language, cost estimated, email addresses, etc),
- the range of values taken by each attribute, i.e., minimum/maximum values (the R functions `min`, `max` and `range` process numeric or character vectors),
- any unexpected or missing values,
- duplicate or very similar rows or columns.

One call to `read.csv` will read the entire contents of a text file into a data frame (what R calls a structure or record type). The file is assumed to contain rows of delimited values (there is an option to change the default delimiter); spurious characters or missing values can cause values to appear in the wrong column (`<`data_preparation`>` provides some suggestions for finding and correcting problems such as this). The `foreign` package contains functions for reading data stored in a variety of proprietary binary forms.

Having read the file into a variable the following functions are useful for exploring what has been read (unless the dataset is small enough to be displayed on a screen in its entirety):

- `str` returns type information about its argument (most usefully the names, types and first few values of each column in a data frame),
- `NROW` and `NCOL` return the number of rows and columns in a vector or data frame (`nrow` and `ncol` are also available but do not behave sensibly when the argument is a vector),
- `head` and `tail` print six rows from the start/end of their argument respectively,
- `table` prints a count of the number of occurrences of each value in its argument, e.g., a particular column of a `data.frame` (by default NAs are not included).

When one or two columns are of specific interest a call to `plot` can be used to quickly visualize the specific data of interest.

For instance, limits imposed by a system can have a significant impact on the usage characteristics of that system. The upper plot in Figure 8.3 shows a very noticeable change in the distribution pattern of C source line length (i.e., number of characters on a line). This change occurs at around the line length commonly supported by non-GUI non-flat screen terminals (these measurements were of C source that is over 10 years old, i.e., before flat screen monitors became available) with wider modern displays and GUI editors the distribution of line lengths seen in more recently written code may be different.

The impact of external limits is not always immediately obvious. The number of tokens per line (lower plot) does not have any dramatic visible changes of behavior, but knowledge that average token length is around 3-4 characters provides a possible explanation for the slight change in the downward slope of the pluses just visible at around 25 tokens.

When a sample contains many variables, plotting one pair of variables at a time is an inefficient use of time. If passed a data frame containing three or more columns, `plot` creates nested plots showing the relationship between every pair of columns. Figure 8.4 shows four sets of related measurements; some measurement pairs appear to have a roughly linear relationship, while no obvious visual pattern is apparent for other pairs.

```
work=read.csv(paste0(ESEUR_dir, "communicating/pub-fs-fp.csv.xz"),
              as.is=TRUE)

# -1 removes the first column
plot(work[, -1], col=point_col, cex.labels=1.2)
```

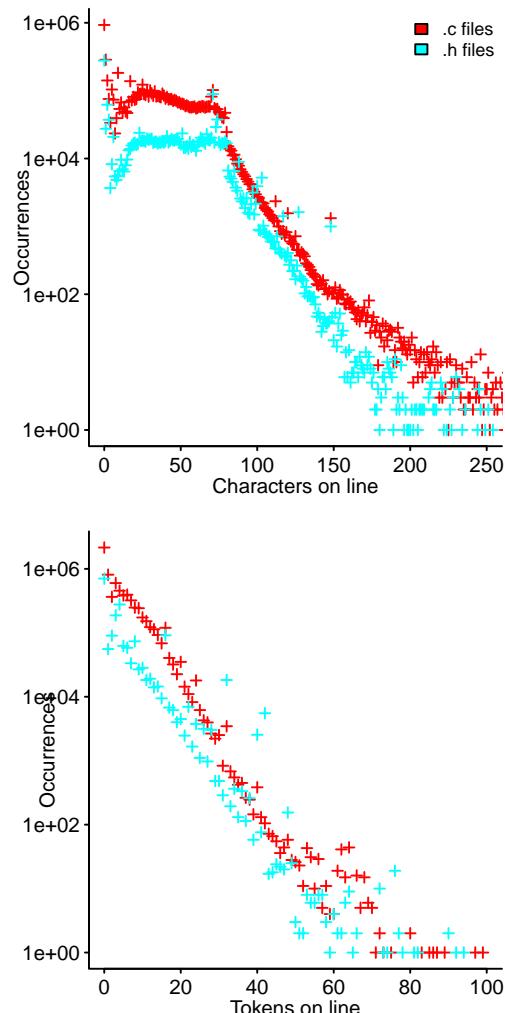


Figure 8.3: Total number of lines of C code, in .c and .h files, having a given length, i.e., containing a given number of characters (upper) and tokens (lower). Data from Jones.<sup>699</sup> code

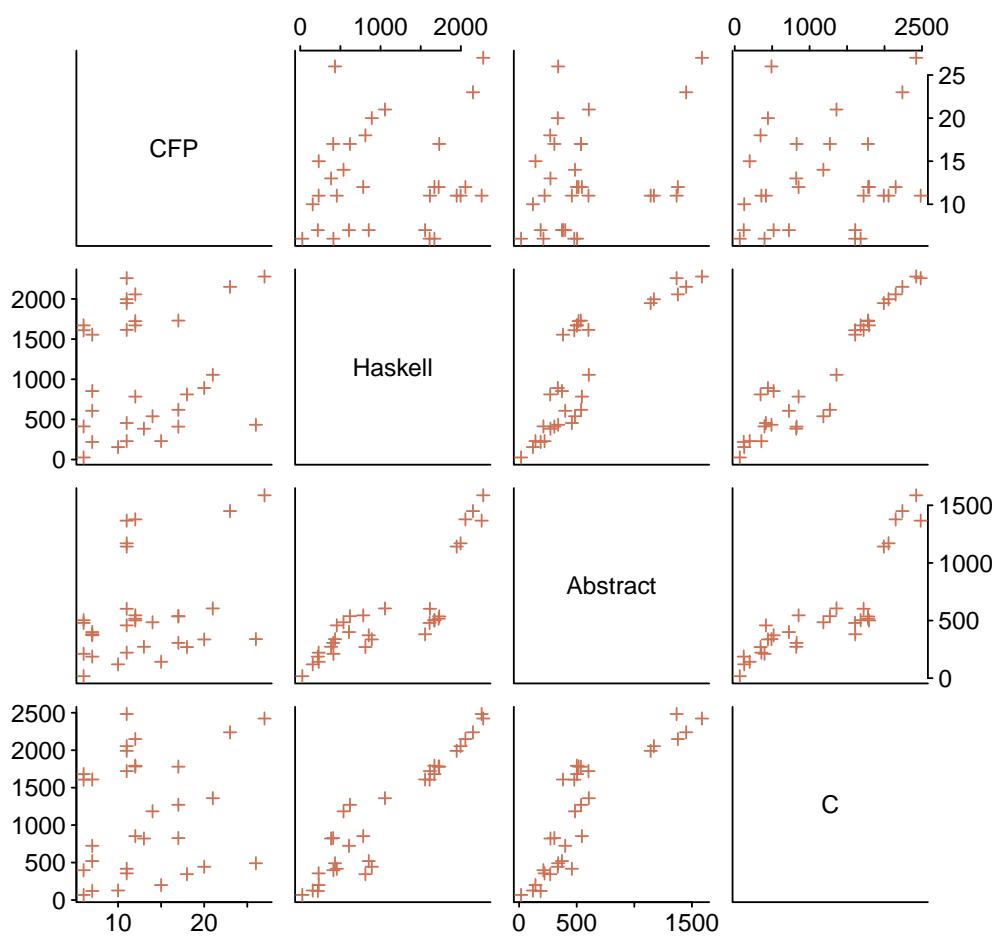


Figure 8.4: Various measurements of work performed implementing the same functionality, number of lines of Haskell and C implementing functionality, CFP (COSMIC function points; based on user manual) and length of formal specification. Data kindly provided by Staples.<sup>1300</sup> code

The `pairs` function supports a variety of options for producing more tailored visualizations of pairs of columns. Separating out and highlighting subsets of a sample (known as *stratifying*) can highlight interesting differences and similarities. Figure 8.5 separates out measurements of Ada and Fortran projects. The lines come from fitting points using loess, a simple regression modeling technique (see below and covered in more detail in a later chapter).

```

panel.language=function(x, y, language)
{
  par(cex.axis=1.1)
  Ada=(language == "Ada")
  points(x[Ada], y[Ada], col=pal_col[2])
  lines(loess.smooth(x[Ada], y[Ada], span=0.7), col=pal_col[2])

  Fortran=(language != "Ada")
  points(x[Fortran], y[Fortran], col=pal_col[1])
  lines(loess.smooth(x[Fortran], y[Fortran], span=0.7), col=pal_col[1])
}

# rows 28 and 30 are zero, and we only want columns 16:19
pairs(log(nasa[-c(28, 30), 16:19]), cex.labels=1.3,
      panel=panel.language, language=nasa$language)

```

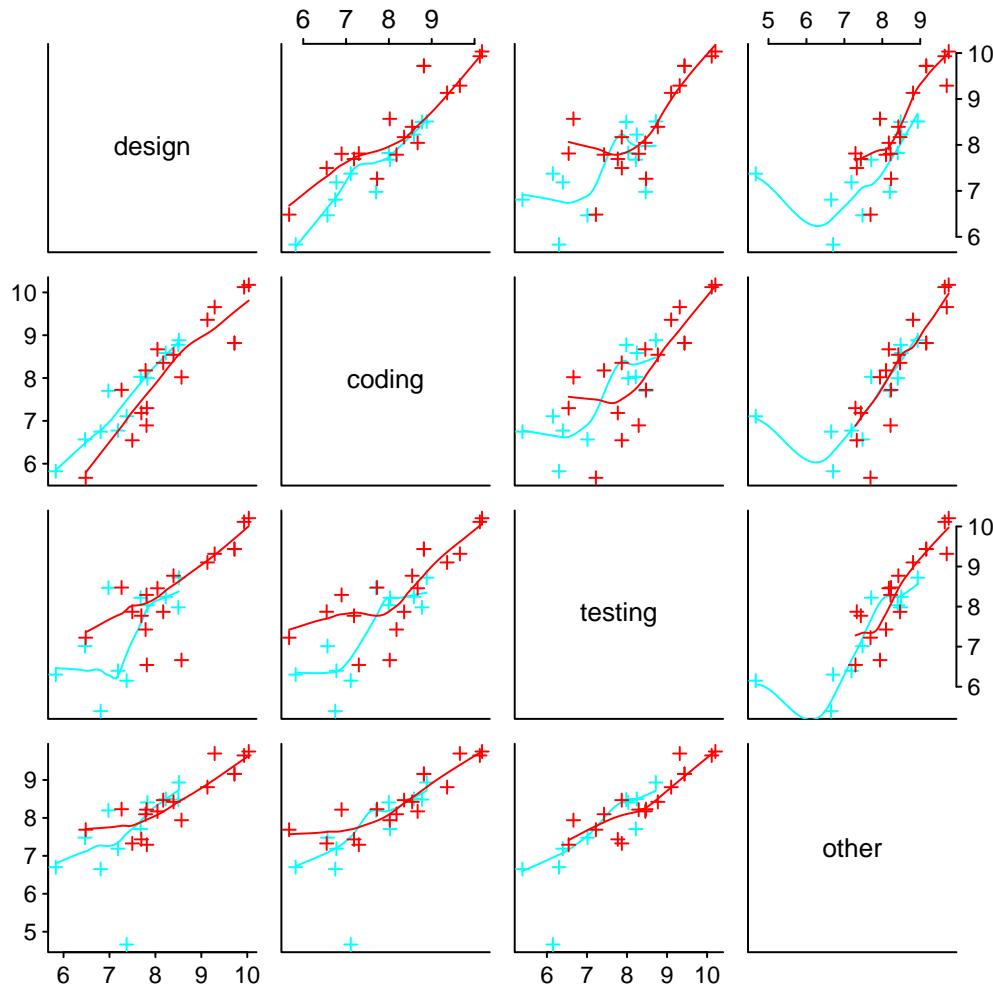


Figure 8.5: Effort, in hours (log scale), spent in various development phases of projects written in Ada (blue) and Fortran (red). Data from Waligora et al.<sup>1420</sup> [code](#)

The default pairs plot contains redundant information; it is possible to display different information in the upper and lower halves of the plot, and along the diagonal. Figure 8.6 shows expert and novice performance (time taken to complete various tasks and final test coverage) in a test driven development task, with a boxplot along the diagonal and correlation between each pair of attributes, for the two kinds of subjects, in the lower half of the plot. This plot, which primarily uses the default values for its visual appearance, would need more work before being presented to customers.

```

panel_user=function(x, y, user)
{
expert=(user == "e")
points(x[expert], y[expert], col=pal_col[1])
points(x[!expert], y[!expert], col=pal_col[2])
}

panel_correlation=function(x, y, user)
{
expert=(user == "e")
r_ex=cor(x[expert], y[expert])
r_nov=cor(x[!expert], y[!expert])
txt = paste0("e= ", round(r_ex, 2), "\n", "n= ", round(r_nov, 2))
text(0.1, 0.5, txt, pos=4)
}

panel_boxplot=function(x, user)
{
t=data.frame(x, user)
boxplot(x ~ user, data=t, border=pal_col, add=TRUE)
}

pairs(tdd[, c("duration.min", "changes", "TDD",
           "average.cycle.length", "development.cycle.length",
           "cycle", "line.coverage", "block.coverage")],

```

```
upper.panel=panel_user, lower.panel=panel_correlation,
diag.panel=panel_boxplot, user=tdd$user)
```

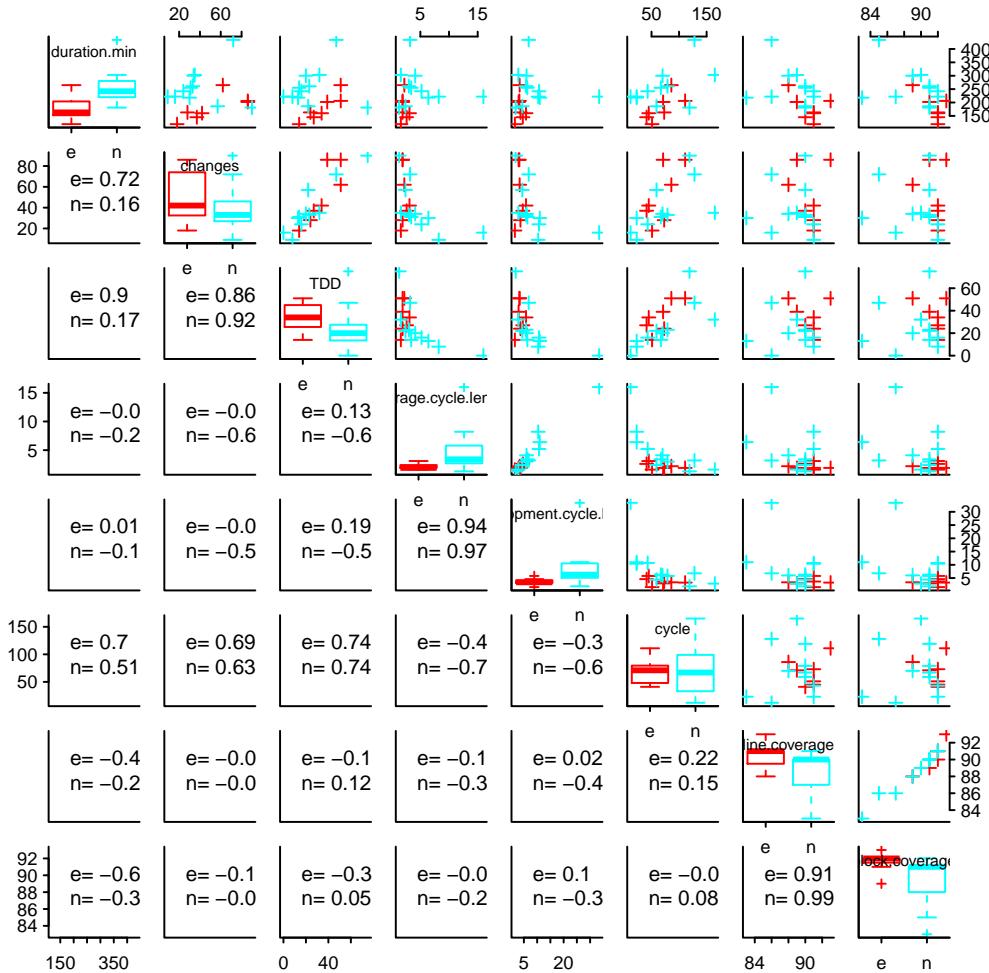


Figure 8.6: Performance of experts (e) and novices (n) in a test driven development experiment. Data from Muller et al.<sup>978</sup> code

The `splom` function in the `lattice` package allows more complex pair-wise plots to be created.

As the number of columns increases, the amount of detail visible in a `pairs` plot decreases. The `plotcorr` function in the `ellipse` package produces a visualization of the correlation between the values in pairs of columns and because correlation is a single value (the extent to which a linear relationship exists) it is possible to create a usable plot containing lots of columns (there are 27 in Figure 8.7). The correlation controls the color and eccentricity of the ellipse, with blue/right slant denoting a positive correlation and red/left slant a negative one.

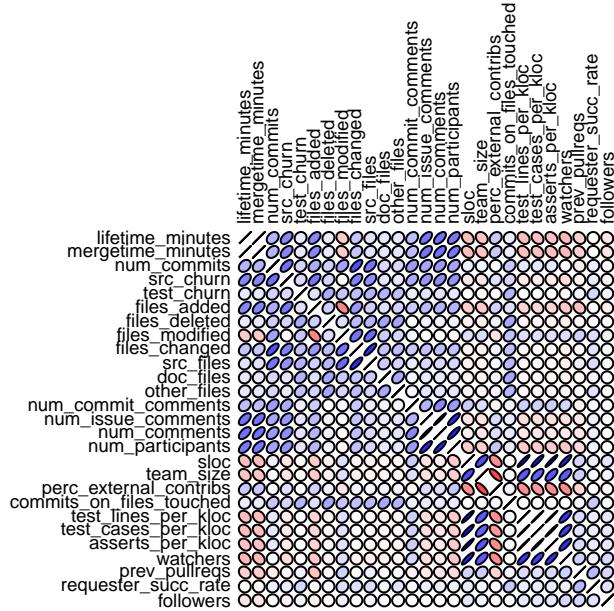


Figure 8.7: Correlations between pairs of attributes of 12,799 Github pull requests to the Homebrew repo, represented using colored ellipses. Data from Gousios et al.<sup>529</sup> code

```
library(ellipse)

# Cross correlation ellipses
ctab = cor(used, method = "spearman", use="complete.obs")

# Map the range 0..1 to colors
colorfun = colorRamp(c("#ff0000", "white", "#0000ff"), space="rgb")
plotcorr(ctab, col=rgb(colorfun((ctab+1)/2), maxColorValue=255),
         outline = FALSE, cex.lab=1.0)
```

The corrgram package supports a variety of functions for displaying correlation information. Figure 8.8 shows one alternative to displaying the data in Figure 8.7. Having variable names appear along the diagonal creates a compact plot; when many variables are involved this form of display is better suited to situations where variable identity follows a regular pattern.

```
library("corrgram")

corrgram(ctab, upper.panel=panel.pie, lower.panel=panel.shade)
```

In a sample containing multiple variables there will be varying degrees of similarity in the patterns of behavior shared by pairs of variables. Hierarchical clustering is one technique for arranging items such that those closer to each other, based on a user supplied distance metric, are closer to each other in the created hierarchy.

The hclust function is included in the base system and requires the user to handle the details; the varclus function in the Hmisc package provides a high level interface. In the following code as.dist is used to map the cross-correlation matrix returned by cor to a distance, see Figure 8.9::

```
ctab = cor(used, method = "spearman", use="complete.obs")

pull_dist=as.dist((1-ctab)^2)
t=hclust(pull_dist)
plot(t, main="", sub="", xlab="Pull related variables", ylab="Height\n")
```

## 8.2.2 Guiding the eye through the data

When looking at a plot of a collection of points, it is not always possible to reliably estimate the path of line through them (according to some goodness of fit criteria). One way to quickly get a rough idea of the likely trajectory of such a line is to use the loess.smooth function (loess is discussed later). Including such a fitted line in a plot is a way of visually showing that expectations of a pattern of behavior is being followed (or not). Figure 8.10 shows a loess

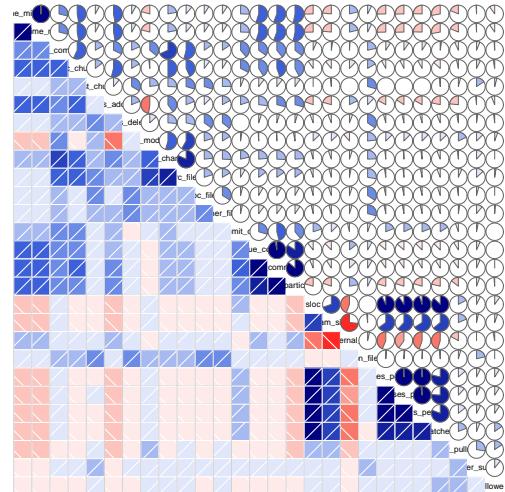


Figure 8.8: Correlations between pairs of attributes of 12,799 Github pull requests to the Homebrew repo, represented using pie charts and shaded boxes. Data from Gousios et al.<sup>529</sup> code

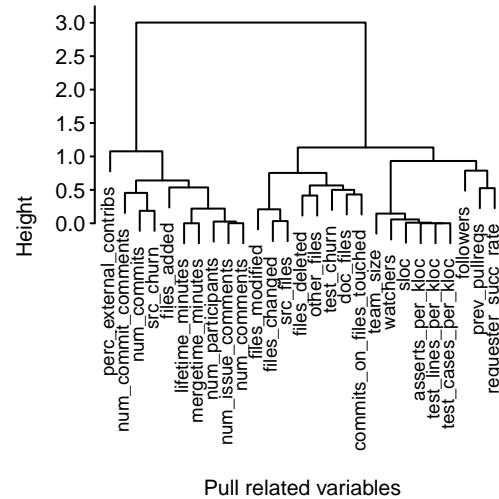


Figure 8.9: Hierarchical cluster of correlation between pairs of attributes of 12,799 Github pull requests to the Homebrew repo. Data from Gousios et al.<sup>529</sup> code

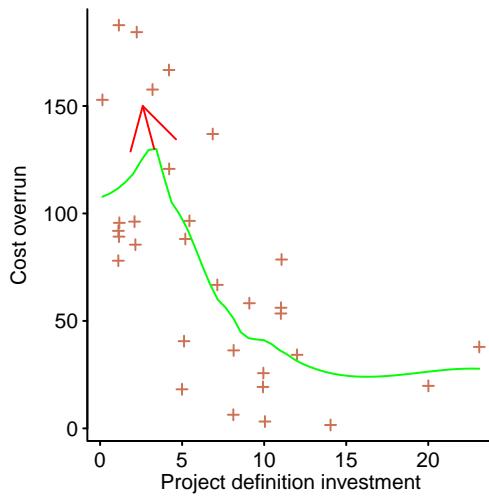


Figure 8.10: Effort invested in project definition (as percentage of original estimate) against cost overrun (as percentage of original estimate). Data extracted from Gruhl.<sup>559</sup> code

fit (green) to NASA data<sup>559</sup> on cost overruns for various space probes against effort invested in upfront project definition; the upward arrow shows the continuing direction of the line seen in the original plot created by NASA users of this data (who are promoting a message that less investment is always bad).

There are a variety of techniques for calculating a smooth line that is visually less noisy than drawing a line through all the points. Splines are invariably suggested in any discussion of fitting a smooth curve to an arbitrary set of points; the `smooth.spline` function will fit splines to a series of points and return the x/y coordinates of the fitted curve.

Splines originated as a method for connecting a sequence of points by a visually attractive smooth curve, not as a method of fitting a curve that minimises the error in some measurement. LOESS is a regression modeling technique for fitting a smooth curve that minimises the error between the points and the fitted curve; the `loess.smooth` function fits a loess model to the points and return the x/y coordinates of the fitted curve.

Both splines and loess can be badly behaved when asked to fit points that include extreme outliers or have regions that are sparsely populated with data. The running median (e.g., `median(x[1:k])`, `median(x[(1+1):(k+1)])`, `median(x[(1+2):(k+2)])`) and so on for some k) is a smoothing function that is robust to outliers; the `runmed` function calculates the running median of the points and return these values (the points need to be in sequential order).

Figure 8.11 shows the maximum clock frequency of cpus introduced between 1971 and 2010; the various lines were produced using the values returned by the `smooth.spline`, `loess.smooth` and `runmed` functions. Don't be lulled into a false sense of security by the lines looking very similar, the *smoothing parameter* provided by each function was manually selected to produce a pleasing looking fit in each case; the mathematics behind the functions can produce curves that look very different and the choice of function will depend on the kind of curve required and perhaps be driven by the characteristics of the data.

```
plot(x_vals, y_vals, log="y", col=point_col,
      xlab="Date of cpu introduction",
      ylab="Relative frequency increase\n")
lines(loess.smooth(x_vals, y_vals, span=0.05), col=pal_col[1])

# smooth.spline and runmed don't handle NA
t=!is.na(x_vals) ; x_vals=x_vals[t] ; y_vals=y_vals[t]
t=!is.na(y_vals) ; x_vals=x_vals[t] ; y_vals=y_vals[t]

lines(smooth.spline(x_vals, y_vals, spar=0.7), col=pal_col[2])

t=order(x_vals)
lines(x_vals[t], runmed(y_vals[t], k=9), col=pal_col[3])
```

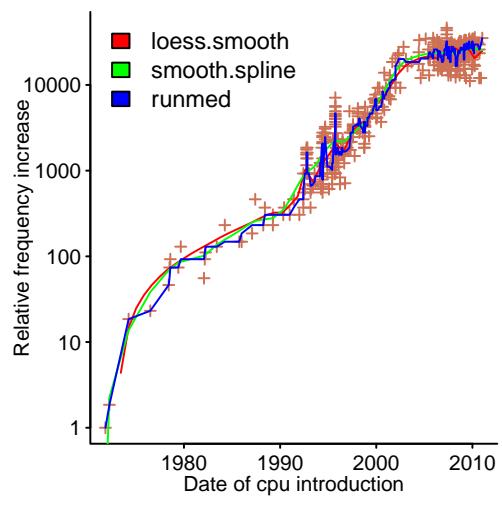


Figure 8.11: Relative clock frequency of cpus when first launched (1970 == 1). Data from Danowitz et al.<sup>312</sup> code

Lines drawn through a sample of measurements values often follow the path specified by some central location metric, e.g., the mean value. In more cases it may be more informative to fit a line such that 25% of the measurements are below/above it, or some other percentage (the fitting process is known as quantile regression). Figure 8.12 is based on 2,183 replies from a survey of FLOSS developers;<sup>1169</sup> two of the questions being the year and age at which those responding first contributed to FLOSS.

If you find yourself writing lots of algorithmic R code during initial data exploration, you are either investing too much effort in one area or you have found what you are looking for and have moved past initial exploration.

### 8.2.3 Smoothing data

Sample values sometimes fluctuate widely around a general trend (the data is said to be *noisy*). Smoothing the data can make it easier to see a pattern in the clutter of measured values. The traditional approach is to divide the range of measurement values into a sequence of fixed width bins and count the number of data points in each bin; the plotted form of this binning process is known as a *histogram*.

Histograms have the advantage of being easy to explain to people who do not have a mathematical background and widespread usage means they are likely to have been encountered

before. Until the general availability of computers, histograms also had the advantage of keeping the human effort needed to smooth data within reasonable bounds.

In the upper plot of Figure 8.13 shows the number of computers having the same SPECint result value, in the lower plot the data has been aggregated into 13 fixed width bins (the number of bins selected by the `hist` function for this data).

The `hist` function accepts a vector of values, automatically selects a bin width and plots a histogram; there are options to change the number of bins and to explicitly specify the location of each bin boundary (which allows variable width bins to be supported, useful when a logarithmic scale is used and bin widths have to follow a geometric progression). The `histogram` package offers automatically support for a wider range of functionality and more optional support that is available in the base system functions. When plotting is not required the `cut` function can be used to divide the range of its argument into intervals and return the bounds of the intervals and the corresponding counts in each interval.

When dealing with measurements that span several orders of magnitude, a log scale is often used. Creating a histogram using a log scale requires the use of bin widths that grow geometrically (coding is needed to get the `hist` function to use variable width bins; the `histogram` package contains built-in support for this functionality) and bin contents has to be expressed as a density (rather than a count). A histogram based on counts, rather than density, can produce misleading results; see [communicating/misc/slash\\_hist.R](#).

The advantage of the binning approach to aggregating data is that it is an easy to perform manual task and for this reason it has a long history. The disadvantages of histograms are: 1) changing the starting value of the first bin can dramatically alter the visual outline of the created histogram and 2) they do not have helpful mathematical properties.

A technique that removes the arbitrariness of histogram bins' starting position is averaging over all starting positions, for a given bin width (known as a *average shifted histogram*); this is exactly the effect achieved using kernel density with a rectangular kernel function.

It often makes sense for the contribution made by each value to be distributed across adjacent measurement points, with closer points getting a larger contribution than those further away. This kind of smoothing calculation is too compute-intensive to be suited to manual implementation, but is trivial when a computer is available.

The distribution of values across close measurement points is known as *kernel density*; histograms are the manual labourer's poor approximation to Kernel density, if a computer is available use the better technique.

The `density` function returns a kernel density estimate (which can be passed to `plot` or `lines`; the following code produced Figure 8.14).

```
plot(density(cint$Result))
```

Density plots also perform well when comparing rapidly fluctuating measurements of related items. Figure 8.15 shows the number of commits of different lengths (in lines of code; upper) to the Linux filesystem code, for various categories of changes and the lower plot shows a density plot of the same data.

The kernel density approach generalizes to more than one dimension; see the `KernSmooth` and `ks` packages.

## 8.2.4 Densely populated measurement points

Some samples contain data having characteristics that are result in plots containing lots of ink and little visual information. Common characteristics of the density of values, such as the following:

- adjacent values on the x-axis having widely different values on the y-values, e.g., the SPECint in the upper plot of Figure 8.13,
- multiple points having the same x/y, visually value appearing as a single value,
- many very similar values that merge into a formless blob when plotted.

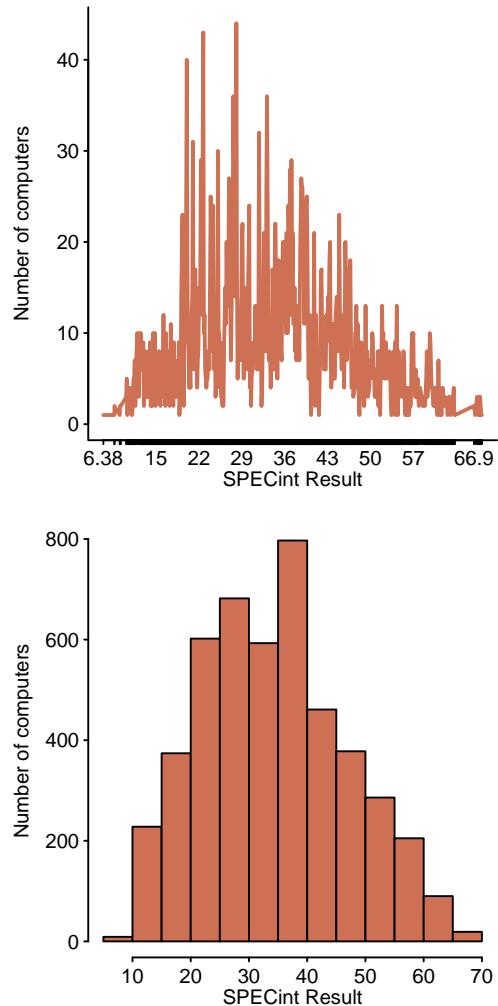


Figure 8.13: SPECint results, summed over all distinct values (upper) and summed within equal width bins (lower). Data from SPEC website.<sup>[1282](#)</sup> [code](#)

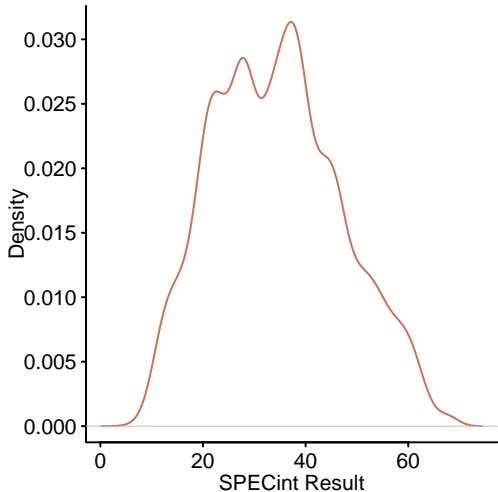
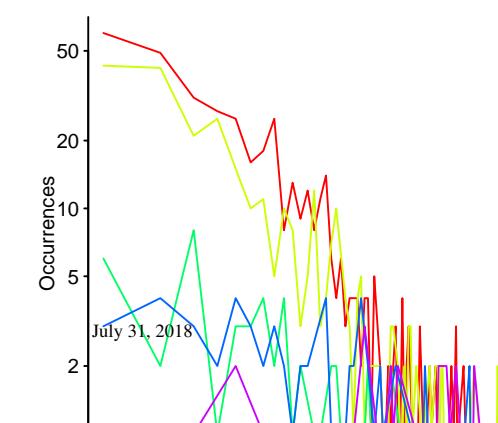


Figure 8.14: Kernel density plot of the number of computers having the same SPECint result. Data from SPEC.<sup>[1282](#)</sup> [code](#)



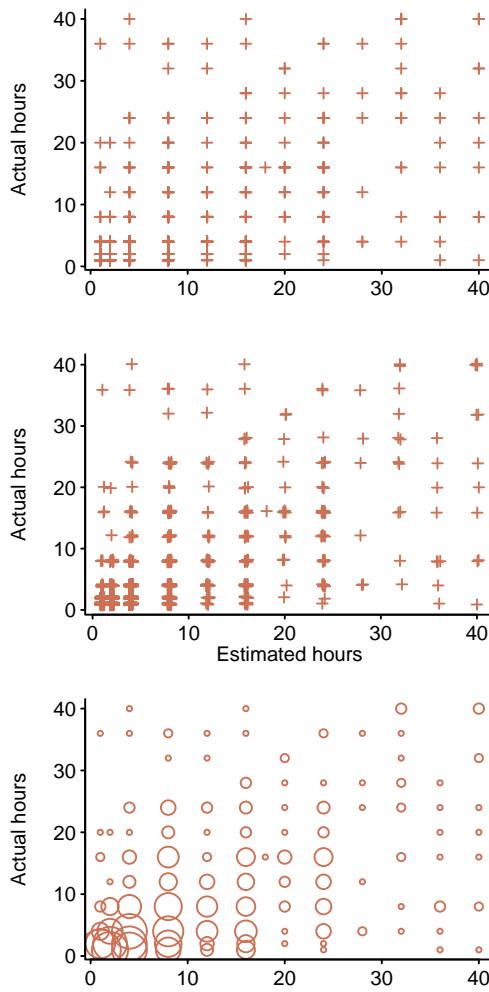


Figure 8.16: Developer estimated effort against actual effort (in hours), for various maintenance tasks, e.g., adaptive, corrective and perfective; upper as-is, middle jittered values and lower size proportional to the log of the number measurements. Data from Hatton.<sup>584</sup> code

A plot of discrete values may give a misleading impression because what appears to be single points may actually be multiple measurements having the same values (see top plot in Figure 8.16). The jitter function returns its argument with a small amount of added random noise; the middle plot of Figure 8.16 shows the effect of jittering the data seen in the upper plot. Another possibility is for the size of the symbol used to vary with the number of measurements at a given point (lower plot of Figure 8.16); as discussed elsewhere, people are poor at estimating the relative area and so size should not be treated as anything more than a rough indicator.

```
plot(maint$est_time, maint$act_time, col=point_col, xlab="", ylab="Actual hours\n")
plot(jitter(maint$est_time), jitter(maint$act_time), col=point_col, xlab="Estimated hours", ylab="Actual hours\n")
library("plyr")
t=ddply(maint, .(est_time, act_time), nrow)
plot(t$est_time, t$act_time, cex=log(1+t$V1), pch=1, col=point_col, xlab="", ylab="Actual hours\n")
```

A different kind of problem occurs when data points are so densely packed together that any patterns which might be present are hidden by the fog (upper plot in Figure 8.17). One technique for uncovering patterns in what appears to be a uniform color surface is to display the density of points. The smoothScatter function calculates a kernel density over the points to produce a color representation (middle plot), or contour lines can be drawn with contour using the 2-D kernel density returned by kde2d (lower plot).

```
plot(udd$age, udd$insts, log="y", col=point_col, xlab="Age (days)", ylab="Installations\n")
# Bug in support for log argument :-( 
smoothScatter(udd$age, log(udd$insts), xlab="Age (days)", ylab="log(Installations)\n")
```

```
library("MASS")
plot(udd$age, udd$insts, log="y", col=point_col, xlab="Age (days)", ylab="Installations\n")
```

```
# There is no log option, so we have to compress/expand ourselves.
d2_den=kde2d(udd$age, log(udd$insts+1e-5), n=50)
contour(d2_den$x, exp(d2_den$y), d2_den$z, nlevels=5, add=TRUE)
```

The hexbin package is available for those who insist on putting values into bins, in this case using hexagonal binning to support two dimensions.

One solution to a high density of point in a plot is to stretch the plot over multiple lines; the xyplot function, in the lattice package, can produce a strip-plot such as the one in Figure 8.18.

```
library("lattice")
library("plyr")
cfl_week=ddply(cfl, .(week),
               function(df) data.frame(num_commits=length(unique(df$commit)),
                                         lines_added=sum(df$added),
                                         lines_deleted=sum(df$removed)))
# Placement of vertical strips is sensitive to the range of values
# on the y-axis, which may have to be compressed, e.g., sqrt(...).
t=xyplot(lines_added ~ week | equal.count(week, 4, overlap=0.1),
          cfl_week, type="l", aspect="xy", strip=FALSE,
          xlab="", ylab="Weekly total",
          scales=list(x=list(relation="sliced", axs="i"),
                      y=list(alternating=FALSE, log=TRUE)))
plot(t)
v 0.10.0
```

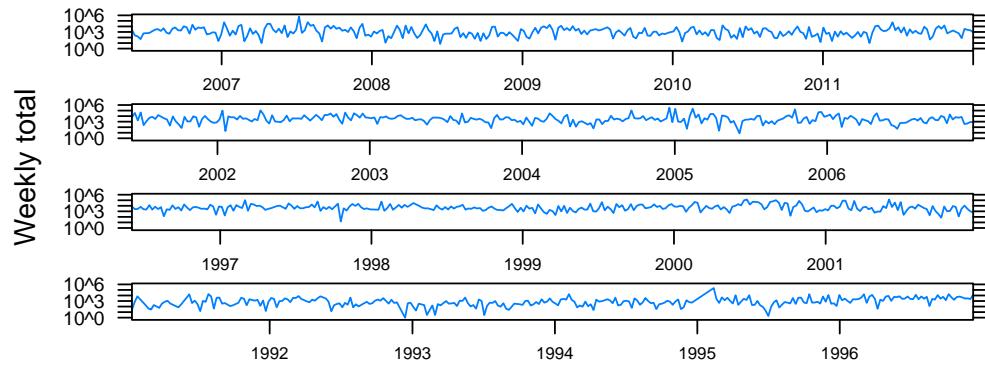


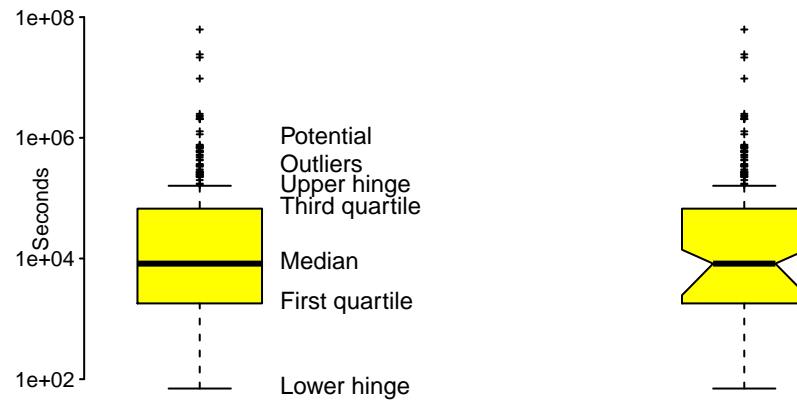
Figure 8.18: Number of lines added to glibc each week.  
Data from González-Barahona et al.<sup>514</sup> [code](#)

## 8.2.5 Visualizing the distribution of values

A *box-and-whiskers* plot (or *boxplot* as it is more generally known) provides a simple visualization of the distribution the values in a sample (see Figure 8.19). The following characteristics are highlighted:

- median, i.e., the point that divides the number of values in half,
- first/third or lower/upper quartile, the 25th/75th percentiles respectively,
- lower/upper hinges, the points at a distance  $\pm 1.5 \cdot IQR$  where  $IQR$  is the interquartile range (the difference between the lower quartile and the upper quartile). The dotted line joining the hinges to the quartile box are the whiskers,
- outliers, all points outside the range of the lower/upper hinge.

The boxplot function produces a boxplot and passing the argument `notch=TRUE` creates a plot that includes a *notch* indicating the 95% confidence interval of the median (right plot in Figure 8.19).



```
box_inf=boxplot(eclipse_rep$min.response.time, log="y",
                 boxwex=0.25, col="yellow", yaxt="n",
                 notch=TRUE, xlim=c(0.9, 1.3), ylab="")
```

The ideas behind the boxplot and kernel density can be combined to create what is known as a *violin plot*. The curve in Figure 8.20 is based on the kernel density of the data points.

The `vioplot` function in the `vioplot` package has limited functionality and for non-trivial displays needs to be used in conjunction with a previous call to, say, the `plot` function. The `beanplot` function in the `beanplot` package supports a lot of functionality, of which producing the envelope of a violin plot is one item.

Figure 8.19: Boxplot of time between a bug in Eclipse being reported and the first response to the report; right plot is notched. Data from Breu et al.<sup>178</sup> [code](#)

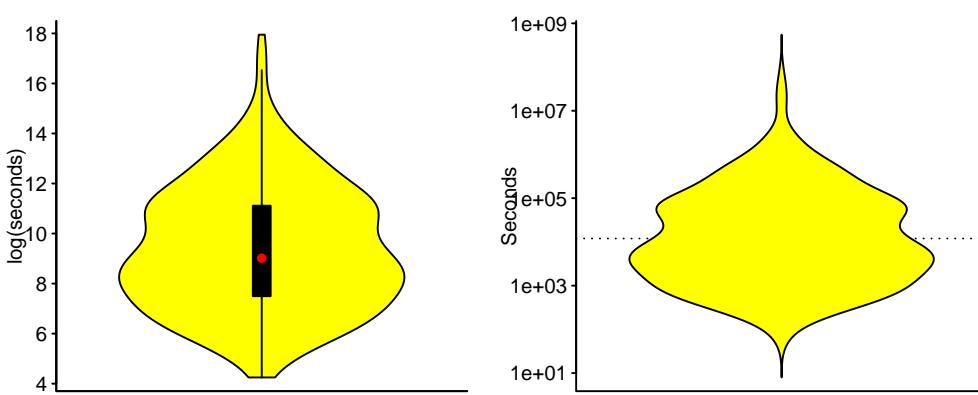


Figure 8.20: Violin plots (left using vioplot, right using beanplot) of time between bug being reported in Eclipse and first response to the report. Data from Breu et al.<sup>178</sup> code

```
library("vioplot")
```

```
# vioplot raises error for attempts to specify axis labels.
# So use plot to control layout and use add=TRUE in vioplot.
# log="y" produces weird results.
plot(x=0.1, xlab="", ylab="log(seconds)", xaxt="n",
      xlim=c(0.2, 1.8),
      ylim=range(log(eclipse_rep$min.response.time)))

vioplot(log(eclipse_rep$min.response.time), col="yellow",
        colMed="red", wex=1.5, add=TRUE)
```

```
library("beanplot")
```

```
beanplot(eclipse_rep$min.response.time, col="yellow", log="y",
         what=c(1, 1, 0, 0), ylab="Seconds\n")
```

The boxplot function supports multiple, related, boxplots in the same plot using formula notation, with the following code producing Figure 8.21 (portions of the notch are offset to avoid obscuring the hinge):

```
boxplot(time ~ group+task, data=gs, notch=TRUE, horizontal=TRUE,
        xlab="Time (minutes)")
```

A bar chart with error bars is regularly used to visually summarise values (sometimes known as *dynamite plots*). A study<sup>289</sup> investigating the effectiveness of various ways of visually summarizing data (including boxplots, violin plots and others) found that when extracting information from bar charts (with or without error bars) subjects did not perform as well as they did when using the other techniques.

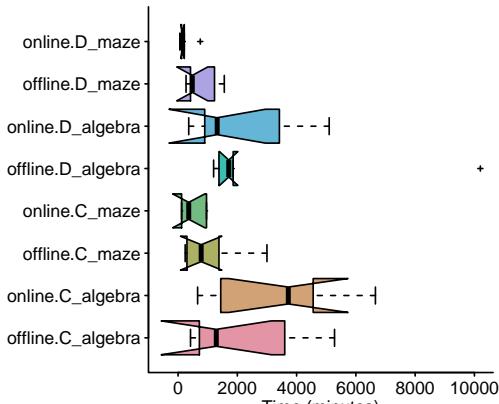


Figure 8.21: Time taken for developers to debug various programs using batch processing or online (i.e., time-sharing) systems. Data kindly provided by Prechelt.<sup>114</sup> code

## 8.2.6 Relationships between items

The relationship between two entities may be the data attribute of interest. The data structure commonly associated with relationships is the graph. The igraph package contains numerous functions for processing graphs.

When displaying graphs containing large numbers of nodes, potentially useful information in the visual presentation can be swamped by many nodes having relatively few connections. Figure 8.22 is an attempt to show which languages commonly occur, in the same project, with another language, in a sample of 100,000 GitHub projects. The number of projects making use of a given pair of languages is represented using line width and to stop the plot being an amorphous blob the color and transparency of lines also changes with number of occurrences.

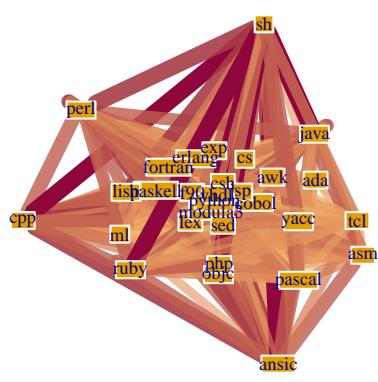


Figure 8.22: Pairs of languages used together in the same GitHub project with connecting line width, color and transparency related to number of occurrences. Data kindly supplied by Bissyande.<sup>145</sup> code

Perhaps those nodes having relatively few connections are the ones of interest. The Microsoft Server protocol specifications<sup>1467</sup> contain over 16 thousand pages across 130 documents (the client specification documents are also numerous). The upper plot in Figure 8.23 shows dependencies between the documents (based on cross-document references in the 2009 release<sup>1467</sup>). The lower plot shows the dependencies after excluding the 18 most often referenced documents.

```

library("igraph")
library("sna")

interest_gr=graph.adjacency(interest, mode="directed")

# V(interest_gr)[names(in_deg)]$size=3+in_deg^0.7
V(interest_gr)$size=1
V(interest_gr)$label.color="red"
V(interest_gr)$label.cex=0.75
E(interest_gr)$arrow.size=0.2

plot(interest_gr)

```

Depending on the question being asked, either the identity of the most frequently referenced documents or the dependencies between those remaining after these have been removed may be included in the story communicated.

It is possible to use R to draw presentable graphs, however, if your primary interest is drawing visually attractive graphs containing lots of information, then there other systems that may be easier to use (e.g., GraphViz<sup>539</sup>). Yes, an R interface to these systems may be available, but if statistical analysis is not the primary purpose, why is R being used?

Alluvial plots are a technique for visualizing the flow between connected entities. Figure 8.24 shows factors used to prioritize the application of Github pull requests and the relative orders in which they appear in a dataset of pull requests.<sup>530</sup>

## 8.2.7 3-dimensions

Three dimensions is only one more than two dimensions and various techniques for enhancing a flat surface to display information about one more dimension (i.e., measurements of a new attribute) are available.

Heatmaps are a technique that use of color to display information about a third quantity within a 2-D plot. Figure 8.25 shows the L3 cache bandwidth of an Intel Sandy Bridge processor when running at various clock frequencies and using various combinations of cores.

Both the heatmap function in the base system and heatmap.2 function in the gplots package clusters the rows/columns and display a dendrogram; various arguments have to be set to switch off this default behavior, with heatmap doing its best to make life difficult including not coexisting with other plots in the same image; heatmap.2 is more reasonable.

The levelplot function in the lattice package provides straightforward functionality for producing heat maps and it is used to produce all the heatmaps in this book.

```

library("lattice")

t=levelplot(L3_band,
            col.regions=rainbow(100, end=0.9),
            xlab="Clock frequency (Mhz)", ylab="Cores used",
            scales=list(x=list(cex=0.70, rot=35),
                        y=list(cex=0.65)),
            panel=function(...)
{
  panel.levelplot(...)
  panel.text(1:11, rep(1:8, each=11),
             L3_band, cex=0.55)
})

plot(t, panel.height=list(3.8, "cm"), panel.width=list(6.2, "cm"))

```

A contour plot can be used for visualizing the relationship between a response variable and two explanatory variables; the contour function is part of the base system.

A study by Thereska, Doebel, Zheng and Nobel<sup>1350</sup> measured the performance of various applications running on a variety of desktop computers; the cpu speed and memory capacity of the computer hosting each of the 4,924,467 user sessions was recorded. The contours in Figure 8.26 are derived from the number of user sessions measured on a computer having a given processor speed and memory capacity.

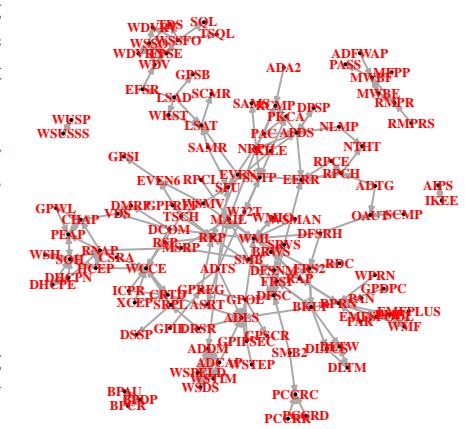
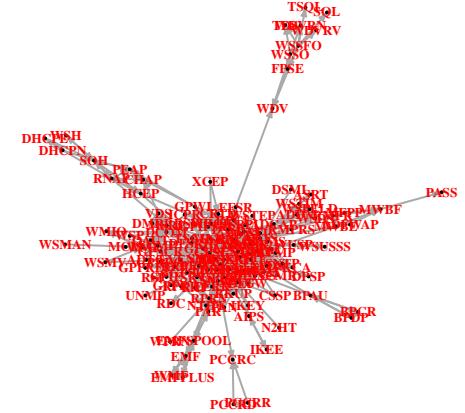


Figure 8.23: References from one document to another in the Microsoft Server Protocol specifications. Data extracted by the author from the 2009 document release. [1467 code](#)

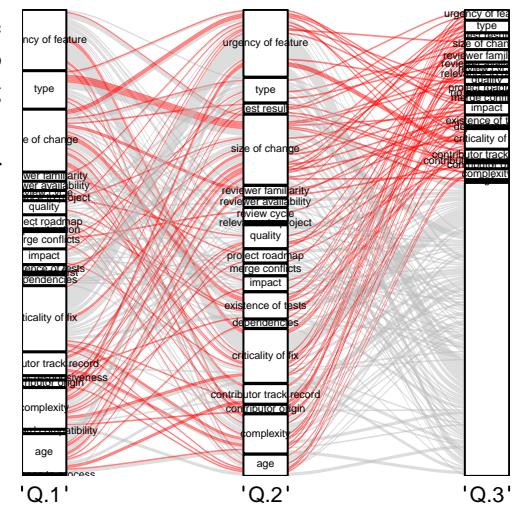
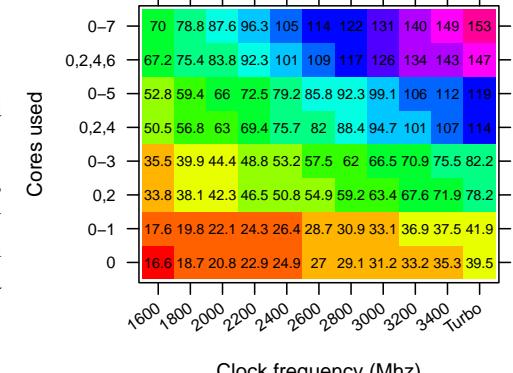


Figure 8.24: Alluvial plot of relative prioritization order of selection and application of GitHub pull requests. Data from Gousios et al.<sup>530</sup> [code](#)



July 31, 2018 Clock frequency (Mhz)

Figure 8.25: Intel Sandy Bridge L3 cache bandwidth in GB/s at various clock frequencies and using combinations of cores (0.3 denotes cores zero through three, 0.2-4 de-

```

library("plyr")

Um=unique(memcpu$MemorySize)
M_map=mapvalues(memcpu$MemorySize, from=Um, to=rank(Um))

Us=unique(memcpu$ProcSpeed)
S_map=mapvalues(memcpu$ProcSpeed, from=Us, to=rank(Us))

cnt_mat=matrix(data=0, nrow=length(Us), ncol=length(Um))

cnt_mat[cbind(S_map, M_map)]=log(memcpu$Session_Count)

contour(x=seq(min(Us)/max(Us), 1, length.out=length(Us)),
         y=seq(min(Um)/max(Um), 1, length.out=length(Um)),
         z=cnt_mat, col=pal_col, nlevels=10, axes=FALSE,
         xlim=c(min(Us)/max(Us), 1), ylim=c(min(Um)/max(Um), 1),
         xlab="Processor speed (GHz)",
         ylab="Memory size (Mbyte)\n")

axis(1, at=sort(Us)/max(Us), labels=sort(Us))
axis(2, at=sort(Um)/max(Um), labels=sort(Um))

```

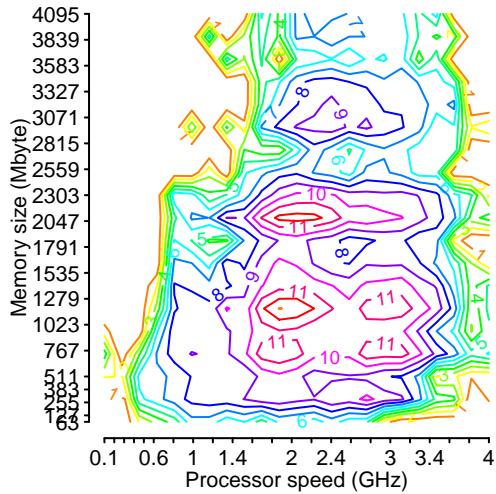


Figure 8.26: Contour plot of the number of sessions executed on a computer having a given processor speed and memory capacity. Data kindly provided by Thereska.<sup>1350</sup>  
code

A variety of functions are available for representing a 3-D plot as a 2-D plot, including the `scatterplot3d` function in the `car`, the `plot3d` function in the `rgl` package.

The `plot.design` function can be used to display the effects of each factor on the mean value of the response variable (see Figure 13.6).

Histograms in 3-dimensions provide more opportunities than histograms in 2-dimensions for looking impressive with little data and misleading viewers. As such, they can be a useful visualization technique.

A study by Hamill and Goseva-Popstojanov<sup>577</sup> investigated the origin of 1,257 faults in 21 large safety critical applications, recording where the fixes were made (e.g., requirements, design, code or supporting files). Figure 8.27 shows a 3-D histogram of root cause/fix location on the x-y axis and a count of occurrences on the z-axis. Color has the effect of enhancing the visual appeal of the plot and makes it easier to locate pairs having similar values, but it is very difficult to obtain detailed information from this plot. Adding numeric values would provide detail, but the real issue is what information is the plot intended to communicate?

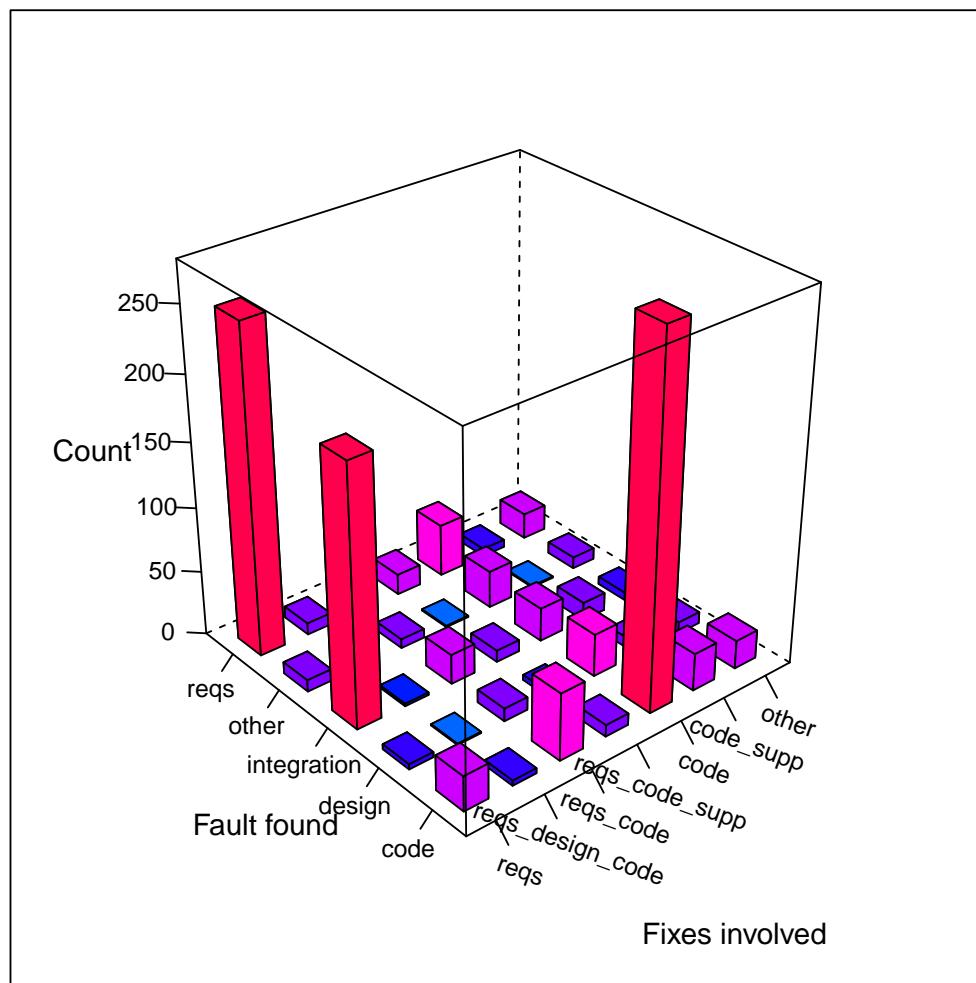


Figure 8.27: Root source of 1,257 faults and where fixes were applied for 21 large safety critical applications. Data from Hamill et al.<sup>577</sup> code

```

library("lattice")
library("latticeExtra")

# log transform pulls out small differences in majority of counts
transform_breaks= exp(do.breaks(range(log(1e-4+STVR_col$occurrences))), 20))
t=cloud(occurrences ~ fix+fault, STVR_col,
        panel.3d.cloud=panel.3dbars,
        xlab="Fixes involved", ylab="Fault found", zlab="Count",
        xbase=0.5, ybase=0.5, aspect=c(1, 1),
        col.facet = level.colors(STVR_col$occurrences,
                                  at = transform_breaks,
                                  col.regions = rainbow),
        scales=list(arrows=FALSE, distance=c(2, 1.1, 1),
                    x=list(rot=-20) # Rotate tick labels
        ))
plot(t)

```

A ternary, or triangle, plot has three axes. The axes are inclined at an angle of 60°, rather than 90°, to each other, and some effort is needed to work out the coordinates of any point. Figure 8.28 shows two ways of labelling a ternary plot (with the three coordinates summing to 100%), with labels appearing at the vertex rather than along the axis and axis scales drawn either perpendicular to the axis or labeled along the axis and within the triangle as a grid. The upper plot shows how lines perpendicular to the appropriate axis are used to find the location of a point (at 10, 35, 55 in this case).

Points appearing close to a vertex have a higher

The closer points are to a vertex the larger the value of the corresponding variable, the closer points are to an axis the smaller the value of the corresponding variable.

Ternary plots are used to visualize compositional data; the compositions and vcd packages include support for creating ternary plots.

In the following code `rcomp` normalises its argument (so that rows sum to 100) using an interval scale and returns an object having class `rcomp` (the compositions has overloaded functions for handing objects of this type):

```
library("compositions")

xyz=c(10, 35, 55)
plot(rcomp(xyz), labels="", col="red", mp=NULL)
ternaryAxis(side=-1:-3, labels=seq(20, 80, by=20), "%"),
      pos=c(0.5,0.5,0.5), col.axis=hcl_col, col.lab=pal_col,
      small=TRUE, aspanel=TRUE,
      Xlab="X", Ylab="Y", Zlab="Z")

lines(rcomp(rbind(xyz, c(10, 45, 45))), col=hcl_col[4])
lines(rcomp(rbind(xyz, c(32, 35, 33))), col=hcl_col[4])
lines(rcomp(rbind(xyz, c(22, 23, 55))), col=hcl_col[4])

plot(rcomp(xyz), labels="", col="red", mp=NULL)

isoPortionLines(col=hcl_col[4])
ternaryAxis(side=0, col.axis=hcl_col, small=TRUE, aspanel=TRUE,
            Xlab="X", Ylab="Y", Zlab="Z")
```

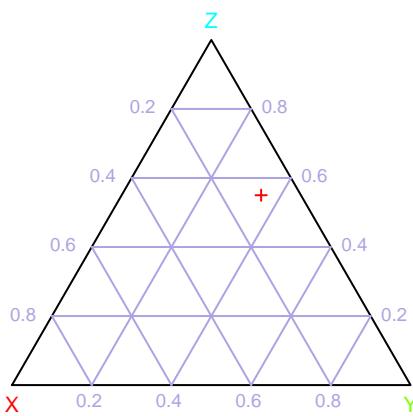


Figure 8.28: Ternary plots drawn with two possible visual aids for estimating the position of a point (red plus at  $x=0.1$ ,  $y=0.35$ ,  $z=0.55$ ); axis names appear on the vertex opposite the axis they denote. [code](#)

## 8.3 Communicating a story

Results from data analysis have no value until they have been reliably communicated to the target audience.<sup>256</sup> Reliably communicating information to other people is difficult; the intended message may be misunderstood or important parts may simply be overlooked by its audience.

No algorithm is known for selecting a method that communicates in a way that is correctly interpreted by its audience, telling the story that is part of the intended message. The main techniques available for presenting numerate information, and how they might be implemented using R, are covered in the rest of this chapter.

There are a wide variety of different ways in which information can be presented, e.g., tables, pie charts, bar charts and scatter plots. Which of these is best at communicating information to readers? The answer from a wide range of studies is that it depends on what information readers are trying to obtain. The following is a brief summary of some research findings:

- graph or table? Studies have found that except for reading-off specific values (and recall of these values later) subjects perform better with line graphs than tables. However, while graphs have better performance when presenting a given perspective (e.g., by selection of the axis), tables may be preferable<sup>170</sup> when wanting to present data in a way that does not favour any one perspective on the data; it boils down to selecting the best cognitive fit,<sup>1401</sup>
- the ability of pie charts to communicate information has been questioned over the years.<sup>300</sup> A study<sup>1284</sup> comparing subject performance using pie charts, a horizontal divided bar chart, a vertical bar charts and a table, found that except when direct magnitude estimation was required pie charts were comparable to bar charts, but for combinations of proportions pie charts were superior,
- adding a third dimension to a graph has been found to slow down reader performance,<sup>613</sup> i.e., subjects take longer to extract information and may be less accurate. The conclusion would appear to be not to use three dimensions when two would do. While subjects have expressed a preference for using 3-D graphs to impress others, no studies have investigated whether they have this effect,
- human judgement of the relative sizes of surface areas has been found not to be based on a linear scale, e.g., a circle of radius two is likely to be around three times larger, rather than four times larger, than a circle of radius one.<sup>i</sup> Encoding information as an area can lead to

<sup>i</sup> Stevens' power law states that psychological magnitude is proportional to the actual magnitude raised to some power. Unlike many other quantities used to express information graphically, for area this power is 0.8 rather than 1 (although values close to 1 have been found for some tasks).

reader communication breakdown; around a quarter of people focus on the areas of each slice in pie charts rather than either the length of the outer arc or the angle at the center of the chart,

- studies by Cleveland are often cited in R related publications: one study<sup>265</sup> asked subjects to make judgements about graphical information encoded in various ways; the results showed that accuracy of subjects' answers varied slightly between encoding methods, ordered from most accurate to least accurate: position along a common scale, positions along nonaligned scales, length/direction/angle, area, volume/curvature and shading/color saturation. Later studies<sup>1284</sup> suggest that things are not so well-defined, with some effects seen being influenced by the structure of the experiments or performance with a particular encoding depending on the task subjects perform.

The thinking behind some layout details used by R's `plot` function are based on experimental work by Cleveland.<sup>264</sup> Although not explicitly stated the aim appears to have been to present data in a workman-like way that avoids the possibility of plotted data values being obscured by plot markings (e.g., tick marks).

The `plot` function is a workhorse for handing the graphical display of data in R; it does a good job of producing a reasonable looking plot from whatever it is passed. Based on this book's implementation goal of using one implementation technique, where-ever possible, the plots in this book were generated using the `plot` function.

The `lattice`<sup>1200</sup> and `ggplot`<sup>233</sup> packages provide alternative world views on the plotting of data; `lattice` is based on the Trellis graphics system<sup>115</sup> from Bell Labs and has an emphasis on multivariate data, while the design of `ggplot` is derived from the work of Wilkinson.<sup>1455ii</sup> Both `lattice` and `ggplot` provide a great deal of control over the created plot through the use of user supplied functions. While `ggplot` is widely used by experienced R developers, its inability to sensibly handle whatever nonsense is thrown at it prevents this package being recommended for casual use. A detailed technical overview of the R graphics subsystems is available in "R Graphics" by Paul Murrell.<sup>982</sup>

Combining data with visual information familiar to readers helps them to extract patterns that mean something to them. Figure 8.29 shows single event upsets (i.e., radiation induced memory faults) experienced by NASA's Orbview-2 spacecraft during one day in 2000. Overlaying the satellite location at the time of the upset on a map of the Earth (using the `map` package) provides context to help readers understand where most upsets occur.

In some cases the intent of a plot may be to communicate that life is complicated. For instance, Figure 8.30 shows an estimate of the market share of Android devices in use in 2015, by brand/company and product name, based on the 682,000 unique devices that downloaded an App from OpenSignal.<sup>1037</sup>

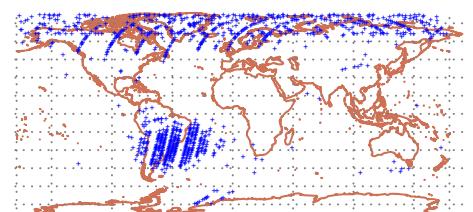


Figure 8.29: Earth relative positions of NASA's Orbview-2 spacecraft when it experienced a single event upset (in blue) on 12 July 2000. Data kindly provided by LaBel.<sup>1099</sup> [code](#)

```
library("treemap")
and_tree=treemap(android, c("brand", "model"), "august2015",
                 title="", palette=pal_col,
                 border.col="white", border.lwds=c(0.5, 0.25))
```

<sup>ii</sup> The title of this book "The Grammar of Graphics" refers to the structure of software written to display graphics rather than the structure of the displayed information.

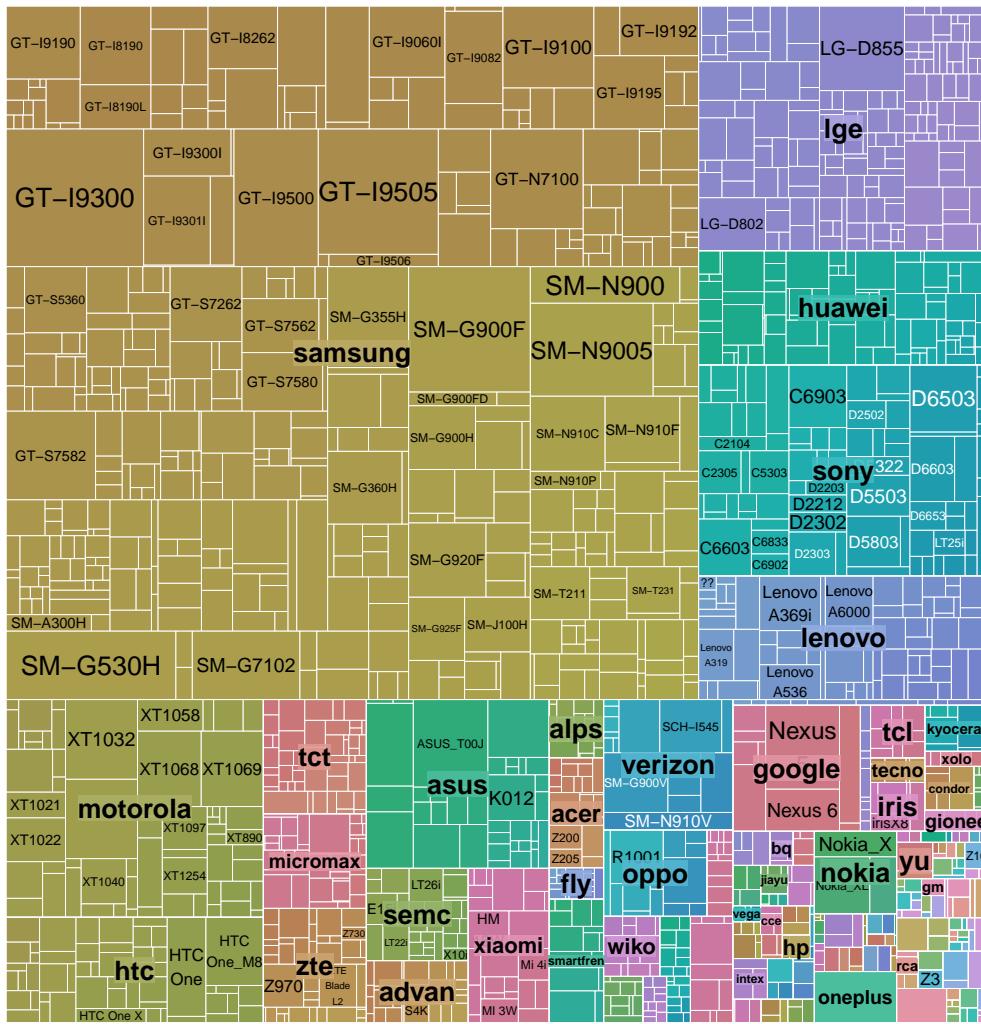


Figure 8.30: Estimated market share of Android devices by brand and product, based on downloads from 682,000 unique devices in 2015. Data from OpenSignal.<sup>1037</sup> [code](#)

### 8.3.1 What kind of story?

The kinds of output from statistical data analysis include the following:

- a description of the data, e.g., its mean and variance, how measurements cluster, an equation that summarises the data. A descriptive model, built from the data, can be used to help gain insights into the system that was measured, for comparing different descriptions (e.g., benchmark results) and for building similar systems (e.g., automatically creating file system contents<sup>13</sup> for benchmarking purposes)
  - a model built to mimic the behavior of a system (as expressed in the measurements made), e.g., a simulator,
  - a predictive model capable of making appropriately accurate predictions using values not in the set of measurements used to build the model. The possible range of prediction values may be within the range of values used to build the model or outside the range of these values, e.g., making predictions about a future time,

A standard reply to any complaints about the adequacy of a model built using data is the adage “All models are wrong but some are useful.”

An example of the different kinds of models that can be built and how their usefulness depends on the problem they are intended to solve is provided by a question involving the usage of local variables in the source code of a function definition.

If the source code contains a total of  $N$  read accesses to variables defined locally within the function, what percentage of variables will be read from once, how many twice and so on (based on a static count of the visible source code, not a dynamic count obtained by executing the function)?

Using data from an analysis of C source<sup>699</sup> we have a description of "what is".

Plotting the data shows that a few variables account for most of the accesses (i.e., read from). After some experimentation the following equation was found to be a good fit to the data; see Figure 8.31:

$$pv = 34.2e^{-0.26acc - 0.0027N}$$

where  $pv$  is the percentage of variables,  $acc$  is the number of read accesses to a given variable and  $N$  is the total number of accesses to all local variables within a function. For example, if a function contains a total of 30 read accesses of local variables the expected percentage of variables accessed twice is:  $34.2e^{-0.26 \cdot 2 - 0.0027 \cdot 20}$ .

Are there other ways of building a model to answer the question asked?

This problem has a form that parallels a model of the growth of new pages and links to existing pages on the world wide web. Each access of a local variable could be thought of as a link to the information contained in that variable. One algorithm that has been found to do a reasonable job of modeling the number of links between web pages is *Preferential attachment*.

With some experimentation an iterative algorithm, based on these ideas, was created that produced a pattern of behavior close to that seen in the data. The algorithm is as follows:

Assume we are automatically generating code for a function and from the start of the function to the current point in the code  $L$  distinct local variables exist (and have been accessed), with each accessed  $R_i$  times ( $i = 1, \dots, L$ ). The following weighted preferential attachment algorithm is used to select the next local variable to access (global variables are ignored in this analysis):

- With probability  $\frac{1}{1+0.5L}$  select a new variable to access,<sup>iii</sup>
- with probability  $1 - \frac{1}{1+0.5L}$  select a variable that has previously been accessed in the function, select an existing variable with probability proportional to  $R + 0.5L$  (where  $R$  is the number of times the variable has previously been read from; e.g., if the total accesses up to this point in the code is 12, a variable that has had four previous read accesses is  $\frac{4+0.5 \cdot 12}{2+0.5 \cdot 12} = \frac{10}{8}$  times as likely to be chosen as one that has had two previous accesses).

The red points in Figure 8.31 were calculated using the above algorithm.

This preferential attachment model provides an insight into local variable usage that is very different from that provided by the fitted exponential equation. neither of them could not be said to be realistic descriptions of the process used by developers when writing code. Both models are descriptions of the end result of the emergent process of writing a function definition. Each model has its own advantages and disadvantages, including the following:

- the fitted equation is fast and simple to calculate, while the output from the iterative model is slow (an average over 1,000 runs in the example code) and requires more work to implement,
- the iterative model automatically generates a possible sequence of accesses (for machine generated source), while a fitted equation does not provide any obvious method of generating a sequence of accesses,
- multiple executions of an iterative model can be used to obtain an estimate of standard deviation, while the equation does not provide a method for estimating this quantity (it may be possible to fit another equation model that provides this information),
- the equation provides an end result way of thinking while the iterative model provides a choice-based way of thinking about variable usage.

A common technique for devising a model for a new problem is to find a very similar problem that has a proven model, and to adapt this existing model to the new problem. A model based on existing practice is often easier to sell to an audience than a completely new model.

Some multiprocessor system have a "shared nothing" architecture, which minimises the sharing of hardware resources. Benchmark performance measurements of such a system under

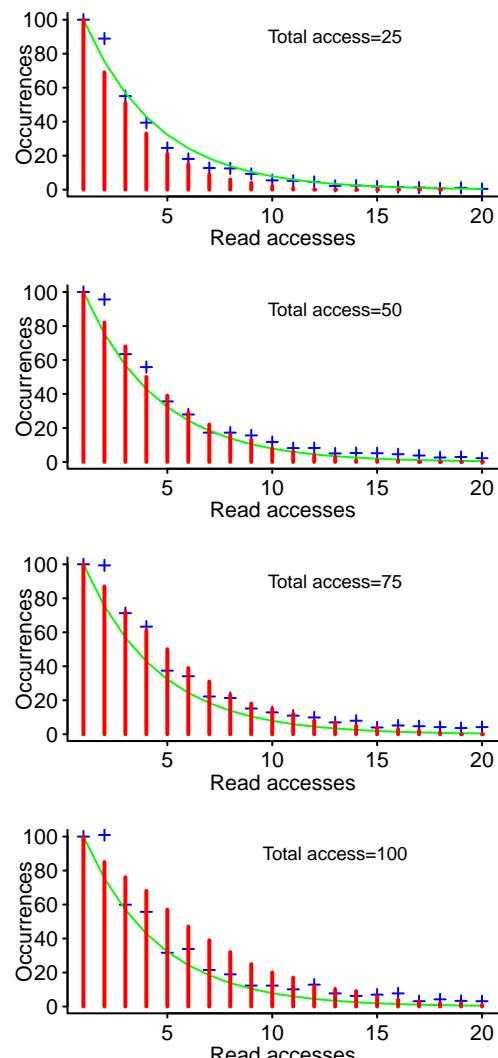


Figure 8.31: Variables having a given number of read accesses, given 25, 50, 75 and 100 total accesses, calculated from running the weighted preferential attachment algorithm (red), the smoothed data (blue) and a fitted exponential (green). [code](#)

<sup>iii</sup> The unweighted preferential attachment algorithm uses a fixed probability to decide whether to access a new variable.

various loads shows that even when tasks can be evenly distributed across all  $X$  processors in the system, performance is rarely  $X$  times faster. What model provides a good explanation of the performance seen?

*Amdahl's law* predicts changes in multiprocessor performance as the number of processors used changes, where the multiprocessor system has a shared hardware architecture. Gunther<sup>564</sup> extended this "law" to cover multiprocessors having "shared nothing" architecture; the adapted model, plus a further adaption, are not good fits to the data (see Figure 8.32).

Gunther<sup>565</sup> then created a model based on queuing theory and ran simulations to model performance (with each job waiting in a queue for time  $t_1$  and executing for time  $t_2$ ). The argument for using queuing theory was that data sharing between different programs can create a resource contention that the "shared nothing" hardware architecture cannot unblock.

Figure 8.32 shows that the queuing model more accurately follows the pattern measured. Given the small amount of data available it would be unwise to attempt further model tuning.

The R language does not contain features designed with simulation in mind<sup>iv</sup>, but like most languages it can be used to solve problems outside of its core domain; see the `simFrame` package.

Finding a good model for the data can sometimes involve many iterations over a long time. For instance, modeling the growth of the size and number of files/directories in a filesystem has a long history, with current models<sup>962</sup> either involving a mixture of two distributions for the equation fitting approach or a generative approach based on simulating the way new files are created from existing files.

Perhaps the most important question to answer when proposing any model is the purpose to which it will be put. A model intended to gain insight might not be of any use in making practical recommendations and a model used to make predictions might not provide any useful insight. For instance, modeling the connection between modifications to files and the introduction of new faults may be used to predict fault rates based on modification history, but this model has limited scope for directly deriving techniques for reducing faults (e.g., reduce faults by reducing file modifications, is of no use when customers want new or modified behavior in the applications they use).

In most cases a great deal of domain knowledge is required to build a model that has the desired level of performance. There is no guarantee that any created model will be sufficiently accurate to be useful for the problem at hand; this is a risk that occurs in all model building exercises. Ideally model building is driven by a theory describing the behavior of the system being modeled. When a theory is complete there is no need for new models, the fact that the creation of a new model is being considered implies that existing models are lacking in some respect.

## 8.4 Technicalities should go unnoticed

The machinery of information presentation should not get in the way of reader's access to that information.

There are many books offering tips, suggests and recommendations for how best to present visual information to readers; the only book highly recommended by your author (it is based on a wide range of empirical research) is "Graph Design for the Eye and Mind" by Stephen Kosslyn.<sup>784</sup> Sometimes multiple plots are used to tell an evolving story, McCloud<sup>911</sup> is a great introduction to this art form.

### 8.4.1 People have color vision

Until the mid 1980s most people used computer terminals that could only display black and white (or green and black). Thirty years later the look-and-feel of computer usage in the mid-1980s still predominates in serious works involving statistical visualization.

<sup>iv</sup> Interfacing to NetLogo<sup>1119</sup>

Figure 8.32: Throughput when running the SPEC SDM91 benchmark on a Sun SPARCcenter 2000 containing 8 CPUs, with the predictions from three fitted queuing models. Data from Gunther.<sup>565</sup> [code](#)

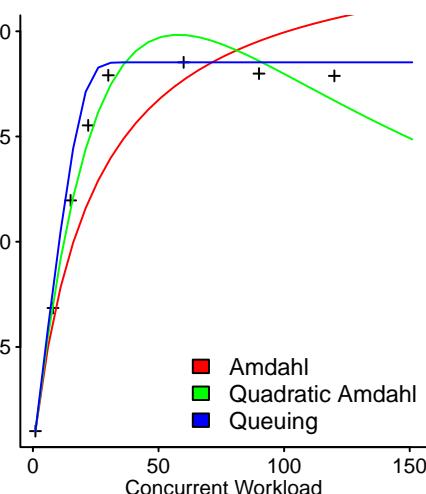
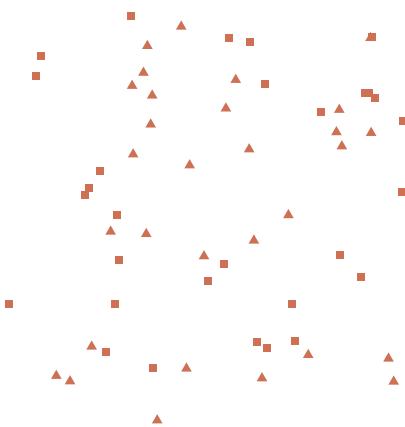
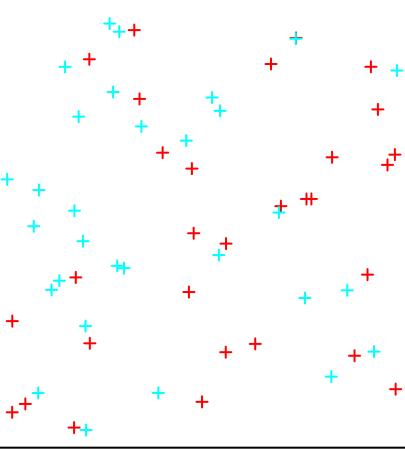


Figure 8.33: Illustration of the difference in cognitive effort needed to locate points differing by shape or color. [code](#)



This book treats color as an essential component of numeric story telling. Color provides an extra dimension that can provide more information within the same space and help the viewer extract information from a plot.<sup>v</sup>

Selecting the most appropriate colors to use requires skill and experience. The `colorspace` and `RColorBrewer` packages both include functions that automatically select a color palette based on the arguments passed;<sup>vi</sup> the `colorspace` package provides a wider range of functionality than `RColorBrewer` and is used to select the colors for the plots in this book. The default usage of the color palette generating functions in this package is to pass the number of colors required in the palette, and assign the returned vector (whose values can be passed as the color argument to plotting functions).

The Hue-Chroma-Luminance (HCL) color space is claimed<sup>1482</sup> to provide a better mapping to the human color perceptual system (hue: dominant wavelength; chroma: colorfulness, intensity compared to gray; and luminance: brightness, amount of gray) than alternative spaces.<sup>vii</sup> The color palettes generated by the `rainbow_hcl` function are considered to be qualitative palettes, that is suitable for depicting different categories; those generated by the `sequential_hcl` function to suitable for coding numerical information that ranges over a given interval, with the `diverge_hcl` function also encoding numerical information but including a neutral value.

The `choose_palette` function provides an interactive, slider based, method for developers to define their own color palettes.

Approximately 10% of men and 1% of women have some form of color blindness. The `dichromat` package provides a way of showing how a plot that contains color would appear to a viewer having some form of color blindness. It does this by making use of experimental data<sup>1409</sup> to simulate the effects of different kinds of color blindness, modifying the requested colors to appear, to normal sighted viewers, like they would to a viewer having the selected kind of color blindness.

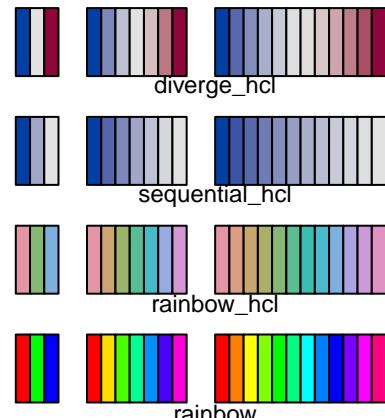


Figure 8.34: The three, seven and twelve color palettes returned by calls to the `diverge_hcl`, `sequential_hcl`, `rainbow_hcl` and `rainbow` functions. [code](#)

## 8.4.2 Color palette selection

Figure 8.35 shows how time varying data involving related items (in this case market share of successive versions of Android) can be displayed in a way that preferentially highlights one aspect of the data; the left plot highlighting individual versions while the right plot shows each version contribution to the overall market share. Bold colors are effective at drawing attention to individual lines, but can be overpowering when there is a large area of color in the plot; the opposite is often the case for pastel colors.

<sup>v</sup> R contains 657 built-in color names (the `colors` function lists them) and also supports hexadecimal RGB literals.

<sup>vi</sup> The selection process is based on theories derived from the use of color in maps,<sup>179</sup> which has a long history.

<sup>vii</sup> Red-Green-Blue (RGB) is a specification based on the display of color on computer screens; Hue-Saturation-Value (HSV) is a transformation of RGB that attempts to map to the human perceptual system and is used by some other software packages.

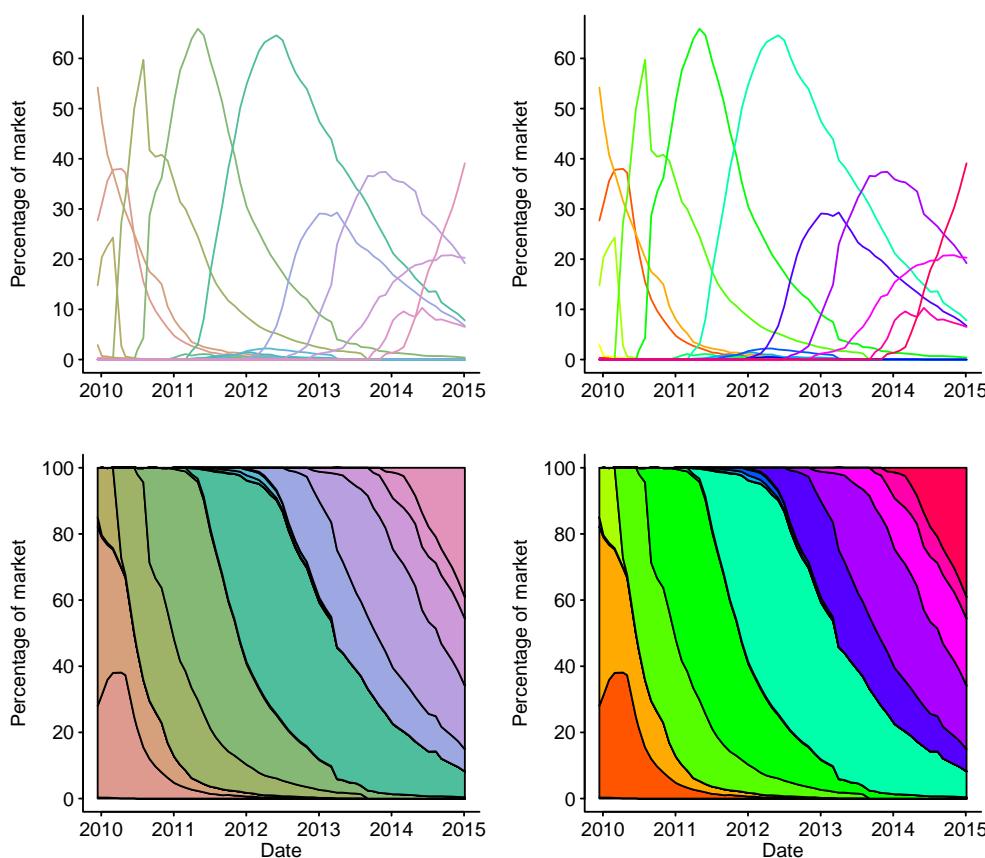


Figure 8.35: Percentage share of the Android market by successive Android releases between 2010 and 2015. Data from Villard.<sup>1410</sup> [code](#)

### 8.4.3 Plot axis: what and how

The choice of plotting axis can have a dramatic impact on the visual perception of the displayed data.

The two commonly used mapping of values to points along an axis are linear and logarithmic. When the range of plotted values spans several orders of magnitude, using a logarithmic axis can produce a more informative visualization; compare the use of linear and log axis in Figure 8.36. Plotting values drawn from an exponential or power law-like distribution using a linear scale often results in many of the points being visually clumped together in a small area of the plot; use of a log scaled axis has the effect of visually expanding these clumped values.

The plot function (and many other plotting functions) automatically select the minimum-/maximum range of each axis based on the range of the data passed; by default 4% is added to each end of the range.

The choice of quantity plotted along each axis is driven by the desire to highlight a relationship between the two quantities; the purpose of a plot is to help viewers appreciate this relationship.

Care needs to be taken to ensure that artificial relationships are not created by the choice of quantity used for an axis. An example of the wasted effort that can occur when the relationship implied by poorly selected quantities is provided by the sorry saga of bug density vs. lines of code.

It was noticed that when bug density (i.e., number of faults divided by lines of code in a function) was plotted against lines of code (in a function), the distribution of points followed a U-shape pattern. Some people proposed that the minimum of this U represented an optimum for the length of a function.<sup>583</sup>

A study by El Emam, Benlarbi, Goel, Melo, Lounis and Rai<sup>381</sup> showed that this U-shape was an artefact generated by the choice of quantities plotted along each axis. Plotting the ratio  $\frac{F}{LOC}$  against  $LOC$ , with  $F$  constant, will produce a tilted U-shape (blue line in Figure 8.37). If the number of faults grows faster than the number of lines of code (which has been found to occur for large line counts) then U-shaped curves such as the red line in Figure 8.37 can occur (a growth rate was picked to illustrate one possibility).

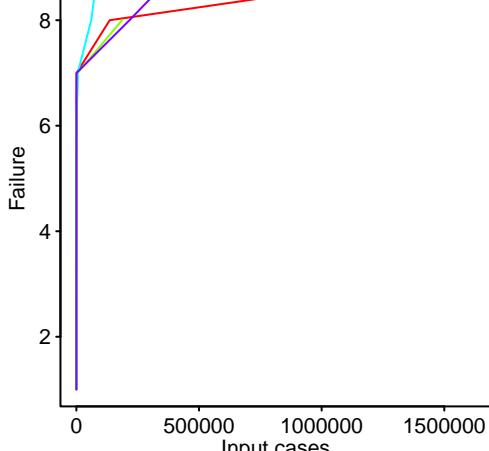


Figure 8.36: Input case on which a failure occurred, for a total of 500,000 inputs; values plotted using a linear (upper) and logarithmic (lower) x-axis. Data from Dunham et al.<sup>366</sup> [code](#)

```
x=1:100 ; inv.x=1/x
plot(x, 3*inv.x, type="l", col=pal_col[1],
      xlab="LOC", ylab="Faults/LOC\n")
lines(x, ((x+50)^3/5e4)*inv.x, col=pal_col[2])
```

The idea suggested by the U-shape pattern in this plot, that there might be an optimal function length, is purely a misinterpretation of the behavior of a ratio quantity plotted against one of the values used in the ratio calculation.

A log transform of an axis can sometimes hide potentially useful information, rather than help reveal it. Figure 8.38 shows Figure 8.3 from a study by Putnam and Myers;<sup>1126</sup> in both cases the x-axis is log transformed. In the right plot the y-axis is linear and there is a visually distinct cluster of measurements, across the top; in the lower plot, where both axes are log transformed, this cluster is visually less prominent.

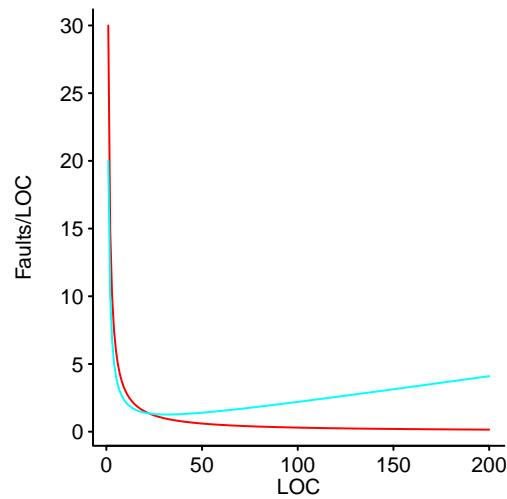


Figure 8.37: Illustration of U-shape created when y-axis values are a ratio calculated from x-axis values. [code](#)

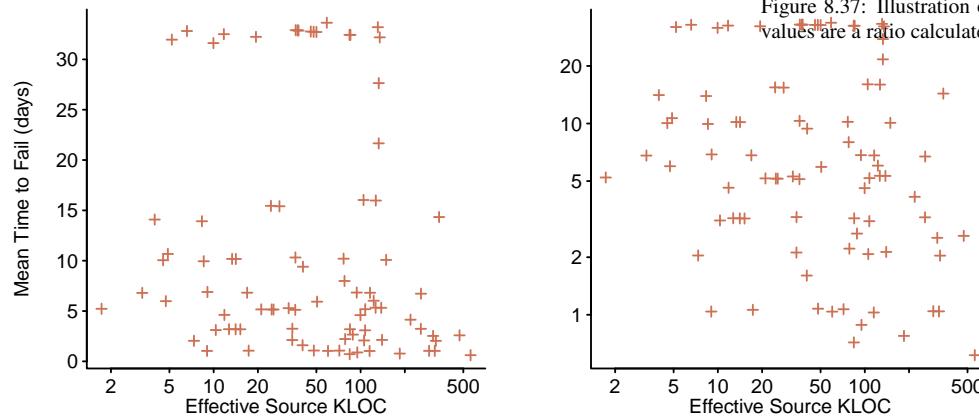


Figure 8.38: Mean time to fail for systems of various sizes (measured in lines of code); linear y-axis left, log y-axis right. Data extracted from Figure 8.3 of Putnam et al.<sup>1126</sup> [code](#)

## 8.5 Communicating numeric values

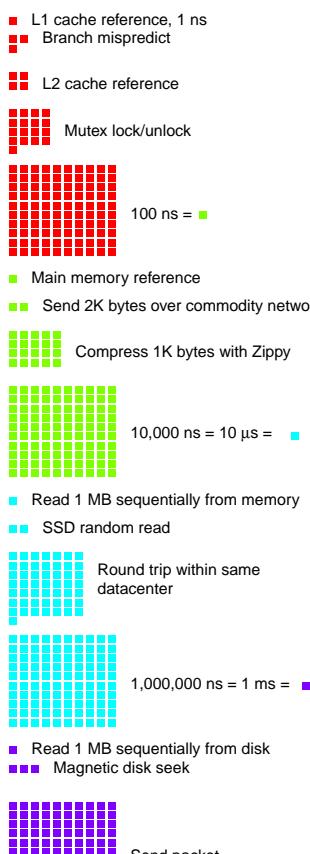
The output from statistical analysis can include visual plots and a small collection of numbers. What is the best way to communicate a story involving a small collection of numbers?

The form of result communication depends on the sophistication of the target audience, with a single value often being preferred for simplicity of selling the result.

A table of numbers covering a very wide range of values (e.g., Table 8.2) can be difficult to interpret quickly unless this is something readers regularly do. An alternative representation separates out the mantissa and exponent, and combines them using area and color, allowing a same/different comparison to be made (see Figure 8.39).

Operation	Approximate runtime
L1 cache reference	1 ns
Branch mispredict	3 ns
L2 cache reference	4 ns
Mutex lock/unlock	17 ns
Main memory reference	100 ns
Send 2K bytes over commodity network	177 ns
Compress 1K bytes with Zippy	2,000 ns
Read 1 MB sequentially: memory	7,000 ns
SSD random read	16,000 ns
Round trip within same datacenter	500,000 ns
Read 1 MB sequentially: magnetic disk	1,000,000 ns
Seek: magnetic disk	3,000,000 ns
Send packet CA→Netherlands→CA	150,000,000 ns

Table 8.2: Numbers Everyone Should Know, circa 2016. Data from Scott.<sup>274</sup>



Confidence bounds are a good way of communicating uncertainty... p-values are often misunderstood...

Multi-way contingency tables using the vcd package...

Replacing numbers with descriptive phrases introduces a whole new level of uncertainty; how do different people interpret descriptions of numeric quantities?<sup>197</sup> emailed...

Some statistical procedures find the values of parameters in an equation that best fit (according to some specified definition of the error) the data. While the numeric values are the output of model building the information being communicated is equation+parameter values, i.e., the final fitted equation should be shown. This suggestion is illustrated in the section on building regression models, as is the use of functions in the sjPlot package plot the parameters of various regression models.

Packages are available for integrating the output from R programs into the workflow of various document preparation systems, for instance, the ascii package provides functions for producing Asciidoc compatible output and the knitr package produces LaTeX output.

Complicated equations can have unexpected behavior. Figure 8.40 shows the result of plotting the following set of equations:

$$-4.7 \leq x \leq 4.7$$

$$y_1 = c(1, -0.7, 0.5) \sqrt{c(1.3, 2, 0.3)^2 - x^2} - c(0.6, 1.5, 1.75)$$

$$y_2 = \frac{0.6 \sqrt{4 - x^2} - 1.5}{1.3 \leq |x|}$$

$$y_3 = c(1, -1, 1, -1, -1) \sqrt{c(0.4, 0.4, 0.1, 0.1, 0.8)^2 - (|x| - c(0.5, 0.5, 0.4, 0.4, 0.3))^2} - c(0.6, 0.6, 0.6, 0.6, 1.5)$$

$$y_4 = \frac{c(0.5, 0.5, 1, 0.75) \tan\left(\frac{\pi}{c(4.5, 4.5)} (|x| - c(1.2, 3, 1.2, 3))\right) + c(-0.1, 3.05, 0, 2.6)}{c(1.2, 0.8, 1.2, 1) \leq |x| \leq c(3, 3, 2.7, 2.7)}$$

$$y_5 = \frac{1.5 \sqrt{x^2 + 0.04} + x^2 - 2.4}{|x| \leq 0.3}$$

$$y_6 = \frac{2||x| - 0.1| + 2||x| - 0.3| - 3.1}{|x| \leq 0.4}$$

$$y_7 = \frac{-0.3 (|x| - c(1.6, 1, 0.4))^2 - c(1.6, 1.9, 2.1)}{c(0.9, 0.7, 0.6) \leq |x| \leq c(2.6, 2.3, 2)}$$

Easy of faking data... [communicating/warp-pts.R](#)

?

### 8.5.1 Percentages vs frequencies

Experiments have found that people are much better at extracting certain kinds of information when it is presented in the form of frequency of occurrence rather than as a percentage<sup>495</sup>... pie charts... inline charts...

Figure 8.39: Alternative representation of numeric values in Table 8.2. Data from Scott.<sup>274</sup> [code](#)

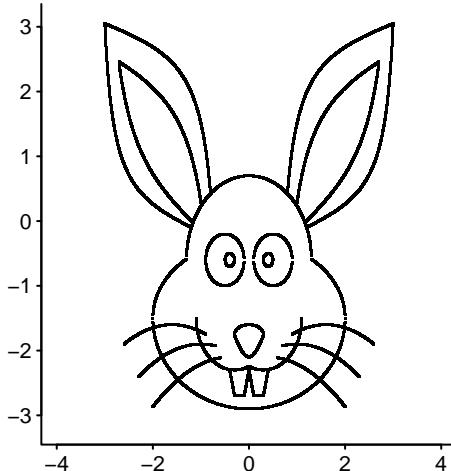


Figure 8.40: What's up doc? Not the fitted model you were expecting. Equations from White.<sup>1442</sup> [code](#)

# Chapter 9

# Probability

## 9.1 Introduction

What are the chances of an event occurring?

Probability is the mathematics involved in answering this question and reasons for being interested in the estimation of probabilities include:

- betting, does a particular case need to be handled, making an insurance decision and for all other decisions and predictions,
- deciding the extent to which an event is surprising. The level of surprise might be used to decide whether something is going wrong or when performing statistical analysis discriminating between hypotheses.

Readers are assumed to have some basic notion of the concept of probability and have encountered the idea of probability in the form of likelihood of an event occurring; classic examples involve calculating the probability of a given combination or sequence of values occurring when flipping a coin or rolling a die, e.g., two heads or rolling two sixes, or the probability of having to make  $N$  flips/rolls before some event occurs.

What is the difference between probability and statistics?

Probability makes inferences about individual events based on the characteristics of the population, while statistics makes inferences about the population based on the characteristics of a sample of the population<sup>i</sup>.

Another way to compare the two is that probability makes use of deductive reasoning while statistics makes use of inferential reasoning.

Probability and statistics are intertwined in that ideas and techniques from probability about individual events may be used when solving problems involving statistics and results about the characteristics of a population obtained from statistical analysis may be used to help solve problems involving probability.

A study by Stewart, Chater and Brown<sup>1314</sup> showed 40 subjects various phrases used to express the likelihood of an event occurring (e.g., "almost impossible" and "quite possible") and asked them to specify the numeric value between 0 and 100 that this phrase suggested... emailed for data.

This book is empirically driven and so primarily makes use of statistical analysis. The following is an example of a problem solved primarily using probability.

The vendor of a static analysis tool wants to add support for detecting a newly discovered potential fault pattern. An occurrence of this pattern in code is not always a fault, what is the upper bound on the probability of generating a false positive that keeps the likelihood that developers will stop using the tool below some limit (say 10%)?<sup>ii</sup>

<sup>i</sup> Statistics could be defined as the study of algorithms for data analysis.

<sup>ii</sup> Experience shows that these false-positives are sufficiently unpopular with developers (they are a source of wasted effort) a developer will often stop using the tool concerned if they are encountered too often. Higher false-positive rates for Tornado warnings result in more deaths and injuries,<sup>1256</sup> through people ignoring the warning.

Answering this question requires knowledge of the mental model used by developers to evaluate analysis tool performance. Two possible mental models include the following (which assumes zero correlation between different warning occurrences and that developers assign the same importance to all warning messages):

- an *economic* developer who tracks the benefit of processing each warning (e.g., false positive warning  $-1$  benefit, else  $+1$  benefit), starting in an initial state of zero benefit this economic developer stops processing warnings if the current sum of benefits ever goes negative.

The Ballot theorem gives the probability that, when sequentially processing warnings, the number of true warnings is always greater than the number of false positive warnings (assuming equal weight is given to both cases, the alternative being more complex to analyse). Let  $C$  be the number of correct warnings and  $F$  the number of false positive warnings and assume  $C > F$ , then the probability is given by:

$$\frac{C - F}{C + F}$$

rewriting in terms of probability of the two kinds of warning we get:

$$C_p - F_p$$

so, for instance, when the false positive rate is 0.25 the probability of a developer processing all the warning generated by a tool is  $0.75 - 0.25 \rightarrow 0.5$ , and does not depend on the total number of warnings.

- an *instant gratification* developer who processes each warning and stops when a sequence of  $N$  consecutive false positive warnings have been encountered. This kind of thinking is analogous to that of the *hot hand in sports* (what psychologists call the clustering illusion).

What is the probability that a sequence of  $N$  consecutive false positive warnings is not encountered?

If the total number of warnings is  $k$  and  $q$  is the probability of a false positive occurring, then the probability of a run of  $N$  consecutive false positive warnings occurring can be calculated using the following recurrence:

$$P(k, q, N) = P(k - 1, q, N) + q^N(1 - q)(1 - P(k - N - 1, q, N))$$

with initial values:

$$P(j, q, N) = 0, \text{ for } j = 0, 1, \dots, N - 1$$

$$P(j, q, N) = q^N, \text{ for } j = N$$

Figure 9.1 shows the probability of not encountering a sequence of three (red) or four (blue) consecutive false positive warnings when processing some total number of warning messages, for various underlying false positive rates (ranging from 0.5 to 0.2).

When dealing with warnings involving complex constructs a developer may be unwilling to put the effort into understanding what is going on and either go along with the what the static analysis tool says, thus underestimating the actual false positive rate, or default to assuming the warning is a false positive, thus overestimating the actual false positive rate.

Finding an equation or technique to use in solving a problem involving probability requires some knowledge of the terminology used in this field. Possible phrases to try in search queries include: birth and death process, coin tossing, colored balls, combination, ergodic, event, fair games, first passage time, generating function, Markov chain, Markov process, occupancy problem, partitions, permutation, random walk, stochastic, trials and urn model.

Finding a closed form solution can be difficult, even when one exists. Simulation using Monte Carlo methods can provide usable estimates of the value of interest.

Sometimes the processes being studied containing so many interacting components that it is not possible to model them analytically; an alternative approach is simulation, discussed in Section 12.4.

### 9.1.1 Useful rules of thumb

If the distribution of the values taken by some attribute, in a population, is not known the following inequalities may be of use as worst case estimates of the probability of various

relationships being true. Both inequalities are distribution independent (the price of this generality is that the bounds are loose).

### Markov inequality

The Markov inequality uses the sample mean,  $X$ , to calculate the maximum probability that  $X$  (which is required to be nonnegative) is larger than some constant. The inequality does not make any assumptions about the sample distribution:

$$P(X \geq k) \leq \frac{\mu}{k}$$

where  $\mu$  is the sample mean.

Example. If a sample of measurements has  $\mu = 10$ , then the probability of the sample containing a value greater than or equal to 20 (i.e., twice the mean) is  $\frac{10}{20}$ .

### Chebychev's inequality

If the standard deviation ( $\sigma$ ) of the sample is known, then Chebychev's inequality can be used to calculate a tighter bound than that given by the Markov inequality, as follows:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

alternatively:

$$P(|X - \mu| \geq k) \leq \frac{\sigma}{k^2}$$

Using the above example the probability of the sample containing a value that differs from the mean by at least 10 is less than or equal to:  $\frac{\sigma}{10^2}$ .

Example: an analysis of the number of mutants needed to estimate test suite adequacy to within a specified error and confidence bounds.<sup>520</sup>

**Fréchet inequalities** Bounds on the union and disjunction of two or more probabilities are given by the Fréchet inequalities, as follows:

Logical conjunction:  $\max(0, P(a_1) + P(a_2) - 1) \leq P(a_1 \& a_2) \leq \min(P(a_1), P(a_2))$

Logical disjunction:  $\max(P(a_1), P(a_2)) \leq P(a_1 \vee a_2) \leq \min(1, P(a_1) + P(a_2))$

**Correlation between three variable pairs** If the correlation between two pairs of three variables is known, say  $r_{12}$  and  $r_{13}$ , the bounds on the correlation of the remaining pair  $r_{23}$  is given by:

$$r_{12}r_{13} - \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)} \leq r_{23} \leq r_{12}r_{13} + \sqrt{(1 - r_{12}^2)(1 - r_{13}^2)}$$

As the number of variables involves increases, the expressions become more complicated.<sup>196</sup>

## 9.1.2 Measurement scales

Mathematically measurement values can be characterised by whether they are discrete or continuous, and the properties of the scale used. Possible scales include the following:

- Discrete
    - nominal scale: each measurement value has an arbitrary number or name. Because the choice of number/name is arbitrary, no ordering relationship exists between different numbers/names. A nominal scale is not a scale in the usual sense of the word.
    - Examples: the numbers on the back of footballers' shirt or the various sales regions in which a product is sold.
  - ordinal scale: each measurement value is a number or name of an item and an ordering relationship exists between the numbers/names. The distance between distinct values need not be the same.
- Example: Classifying faults by their severity, e.g., minor, moderate, serious.
- If a minor fault is considered less important than a moderate fault, and a moderate fault is less important than a serious fault we can deduce that a minor fault is less important than a serious fault.

English	tree	wood	forest
French	abre	bois	forêt
Dutch	boom	hout	bos
German	Baum	Holz	Wald
Danish	træ		skov

Figure 9.2: The relationship between words for tracts of trees in various languages. The interpretation given to words (boundary indicated by the zigzags) in one language may overlap that given in other languages. Adapted from DiMarco et al.<sup>354</sup>

The address of the members of a structure type increases for successive members, but the difference between member addresses is not fixed because different members can have different types.

When names are assigned to entities, there may be cultural differences in the selection process. Figure 9.2 shows how words are assigned to tracts of trees having various areas.

- Continuous

- interval scale: each measurement value is a number and not only does a relative ordering exist but a fixed length interval of the scale denotes the same amount of quantity being measured.

A data point of zero does not indicate the absence of what is being measured.

Example: the start date of some event is an interval scale. If the start date of events A, B and C are known, and difference in start date between events A and B is the same as between events C and D, then it is possible to calculate the start date of event D.

Addition and subtraction can be applied to values on an interval scale but not multiplication or division (e.g., it makes no sense to say that the start date of event A is twice that of event C).

- ratio scale: each measurement assigns a number to an item and this numeric scale preserves: the ordering of items, the size of the interval between items and the ratios between items. It differs from the interval scale in that a measurement of zero denotes the lack of the attribute being measured.

The time difference between two events is a ratio scale.

The kinds of statistical analysis that can be legitimately performed on the values in a sample will depend on the kind of scale used to measure values.

## 9.2 Probability distributions

Probability distributions are mathematical descriptions of the properties of values obtained by following a consistent pattern of behavior, e.g., the flipping a coin pattern of behavior generates one of two results, a fixed probability of either result, with each result being independent of the previous one and a count of the number of heads and tails has a binomial probability distribution.

If a sample of values can be fitted to a known probability distribution, then information about the pattern of behavior that generated them can be inferred from what is known about processes generating values having that particular distribution. For instance, given a list of pairs of numbers, if the ratio formed from each pair (i.e.,  $\frac{a}{a+b}$ ) can be fitted to a binomial distribution, then there is strong evidence that the pairs are counts of a process producing one of two possible values (e.g., heads/tails, yes/no etc) and the probability of producing each value can be calculated from the fitted distribution.

Fitting a distribution to a sample is a step towards understanding the processes that generated the measurements, not an end in itself.

Failure to fit values to a known distribution may mean that more than one distribution is involved, e.g., two different coins are being used and both are biased in some way. Given enough data it is sometimes possible to obtain a reasonable fit that involves two or more distributions.

If there is a reason for believing that the measured processes are driven by particular behaviors, the quality of fit of the predicted probability distribution to the sample can be compared against the quality of fit of other distributions.

If there is no expectation of a particular behavior, then finding an acceptable fit of a probability distribution to the sample values is a starting point for understanding the processes that are driving the measurements observed.

While many probability distributions have been created,<sup>255</sup> only a handful of them are regularly used; R packages tend to support commonly occurring distributions with a few packages supporting a wide range of distributions.<sup>255</sup>

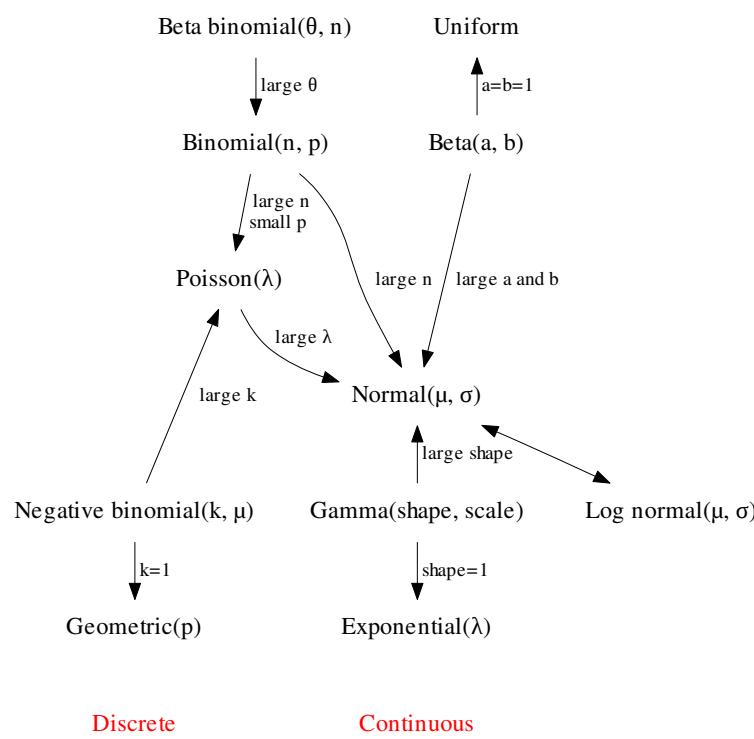


Figure 9.3: Relationships between common discrete and continuous probability distributions.

Every family of probability distributions is completely characterised by a small set of numbers (often one or two) and a formula that the numbers parameterise. For instance everything about a Normal probability distribution can be calculated by plugging values for the mean,  $\mu$ , and standard deviation,  $\sigma$ , into the formula:  $P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$  (this formula is often abbreviated as  $N(\mu, \sigma)$ ). Fitting data to a Normal distribution involves finding appropriate values for  $\mu$  and  $\sigma$ .

In practice a few probability distributions are encountered much more frequently than others. One of these common cases will often fit reasonably well to a wide spectrum of commonly encountered samples and unless there are theoretical reasons for expecting one of the less commonly encountered distributions then there is nothing to be gained by searching through all known distributions to find the one that is the best fit for a sample.

Some of the characteristics of plotted sample values, that may or may not correspond to a known probability distributions include:

- mean value (sometimes called the *arithmetic mean*, *central tendency* or *location value*, the last two terms may also be used to refer to the median): many distributions have a finite mean (some that don't include power laws with an exponent greater than or equal to  $-1$  and the Cauchy distribution),
- scale parameter, variance (standard deviation is the square-root of variance): how spread out the distribution is, a few distributions do not have a finite variance, e.g., power laws with an exponent between zero and  $-2$ ,
- symmetrical/asymmetrical about the mean, the *skew* of a distribution is a measure of how asymmetrical it is; a symmetric distribution has a skew of zero, while a positive skew has a tail pointing towards larger positive values and a negative skew has a tail pointing towards negative values.
- where most of the density resides, i.e., around the mean or in the tails. The *kurtosis* of a distribution is a measure of how spiky the distribution is, possibilities include tall and slim (known as *leptokurtic*; slender-curved), short and flat (known as *platykurtic*) or medium-curved (known as *mesokurtic*; the Normal distribution has a Kurtosis of three).
- number of distinct peaks, known as the *modality* of a distribution; a distribution with one distinct peak is said to be unimodal, two distinct peaks bimodal (such as measurements from two different distributions, e.g., height of men/women),

The `moments` package contains functions for calculating skewness, kurtosis and moment related attributes of a numeric vector.

Probability distributions can be divided into discrete and continuous distributions, with discrete distributions only being defined at specific points (usually integer values). In R functions that involve discrete distributions usually require integer values while functions involving continuous distributions takes floating-point values.

There are various ways of representing a probability distribution and the following are often encountered:

- density function: for discrete distributions, see Figure 9.4, this can be viewed as the probability that  $x$  will have a given value,  $P(x = \text{value})$ ; for continuous distributions, see Figure 9.6, the probability of any particular value occurring is zero, however there is a finite probability of a measurement returning a value within a specified interval.
- cumulative density function: the probability that  $x$  will be less than or equal to a given value  $P(x < \text{value})$ , see Figure 9.5,
- equation: an equation for the probability distribution. For the majority of people this is little more than eye candy, e.g., the equation,  $\frac{\lambda^k e^{-\lambda}}{k!}$ , is very difficult to visualize and is only of use to developers wanting to implement the Poisson distribution.

**Discrete distributions:** commonly encountered discrete distributions include the following (see Figure 9.4):

- Binomial distribution: for a random variable  $X$ ,

1. the process involves a sequence of independent trials,
2. each trial produces two possible outcomes, e.g., heads/tails,
3. the probability of either outcome (say  $p$  for heads) does not change,
  - $X$  counts the number of success (where success might be defined as a head occurring) in  $n$  fixed trials.

The Binomial distribution is completely described by two parameters:  $B(n, p)$ ,

The above process is sometimes described as the process of drawing  $n$  objects from a pool containing a finite number of two kinds of object, where the object is placed back in the pool after it has been drawn (the draws are said to be *with replacement*).

The Hypergeometric distribution is obtained if objects are not returned to the pool once they are drawn (the draws are said to be *without replacement*).

A distribution that takes more than two discrete values is known as a *Multinomial distribution* (again with a fixed probability of each value occurring). The *XNominal* package provides support for multinomial distributions,

- Negative Binomial distribution: this has the same three requirements as the Binomial distribution, and differs in what is counted,

- $X$  counts the number of trials up to and including the  $k^{th}$  success (where success might be defined as a head occurring after a continuous sequence of tails).

Another process that produces values having a Negative Binomial distribution is randomly drawing from a mixture of Poisson distributions, where the mean of the mixture of Poisson distributions has a Gamma distribution,

This distribution is a generalised version of the Geometric distribution (which is based on the probability of observing the first success on the  $n^{th}$  trial).

- Poisson distribution: for a random variable  $X$ ,

1. the process involves independent events,
2. only one event can happen at any time,

- $X$  counts the number of events that occur within a specified time.

The Poisson distribution is completely described by one parameter ( $\lambda$ , the distribution mean):  $P(\lambda)$ ,

The sum of two independent Poisson distributions  $P(\lambda_1)$  and  $P(\lambda_2)$  is the Poisson distribution  $P(\lambda_1 + \lambda_2)$ .

The Binomial and Poisson distributions are related in that as  $n \rightarrow \infty$  and  $p \rightarrow 0$ , then  $B(n, p) \rightarrow P(np)$ , i.e., The Poisson distribution is a limit case of a Binomial distribution having a very low probability of success over a long period.

**Continuous distributions:** commonly encountered continuous distributions include the following (see Figure 9.6):

In all but one case the generating process clusters the values around a single peak.

- Uniform distribution: all values between the lower and upper bounds of the interval have an equal probability of occurring, i.e., no value is more likely to occur than any other. For discrete values between 1 and  $n$  the probability of any value occurring is  $\frac{1}{n}$ .

One process that generates a uniform distribution is a random number generator, such as calling the `runif` function.

- Normal distribution: this can be generated by adding together contributions from many independent processes, a consequence of the Central limit theorem. This distribution crops up with great regularity, it has a mathematical form that is much easier to analytically manipulate than many other distributions resulting in it being widely used before computers reduced the need for analytic solutions to problems. This distribution is described by its mean and variance.

While the Normal is the result of adding contributions from many independent processes, it is not true to say that adding contributions from many different kinds of processes will result in this distribution, similarly for multiplicative contributions and a lognormal distribution. For instance, given the right conditions, adding values drawn from many different Poisson distributions can result in a Negative Binomial distribution, a Geometric distribution and many other distributions,<sup>735</sup>

- Lognormal distribution: the logarithm of a Normal distribution and it can be thought of as being generated by multiplying together the sum of contributions from many independent processes;<sup>961</sup> samples drawn from a Lognormal distribution can produce a straight line, over some of their range, when plotted using log-log axis,

- Exponential distribution: generated by a memoryless process, e.g., when the waiting time for an event to occur is independent of the amount of time that has passed since the last event. This is the continuous form of the Geometric distribution, and like it, is described by a single parameter.

Over some of its range the exponential distribution is visually very similar to a Power law, which has led researchers to incorrectly claim that their sample fits a power law (a fashionable distribution to have one's sample following).

The sum of a reasonably large number of independent exponential distributions has an Erlang distribution, e.g., the interval between incoming calls to a telephone exchange, where the interval between calls from an individual have an exponential distribution,

- Beta distribution: applies to processes where the explanatory variable is restricted to a finite interval, e.g., zero to one. This distribution is defined by two shape parameters.

- Gamma distribution ( $\Gamma$  is the Greek uppercase Gamma, the symbol often used to denote the Gamma function;  $\gamma$  is the lowercase Gamma): used to describe waiting times, e.g., `Gamma(shape=3, scale=2)` is the distribution of the expected waiting time (in some units) for three events to occur given that the average waiting time is 2 time units (yes, the Gamma function differs from most other distribution names in the base system by starting with an uppercase letter).

When `shape=1`, the Gamma distribution reduces to the Exponential distribution.

The Gamma distribution is the continuous equivalent of the Negative binomial distribution.

- Chi-squared distribution (sometimes written using  $\chi$ , the Greek lowercase letter of that name): this is more often encountered in the mathematical analysis of statistics than as a distribution of a sample. A random variable has a chi-squared distribution with  $d$  degrees of freedom if it is produced by a process which generates the values:  $Z_1^2 + Z_2^2 + \dots + Z_d^2$ , where  $Z_i$  are independent random variables having a Normal distribution.

The chi-squared distribution is a special case of the gamma distribution.

- Weibull distribution: this distribution drops out as the solution to various problems in hardware reliability, e.g., time to failure, and is often used as the hazard function in survival analysis. The Exponential and Rayleigh distributions are special cases of the Weibull distribution,

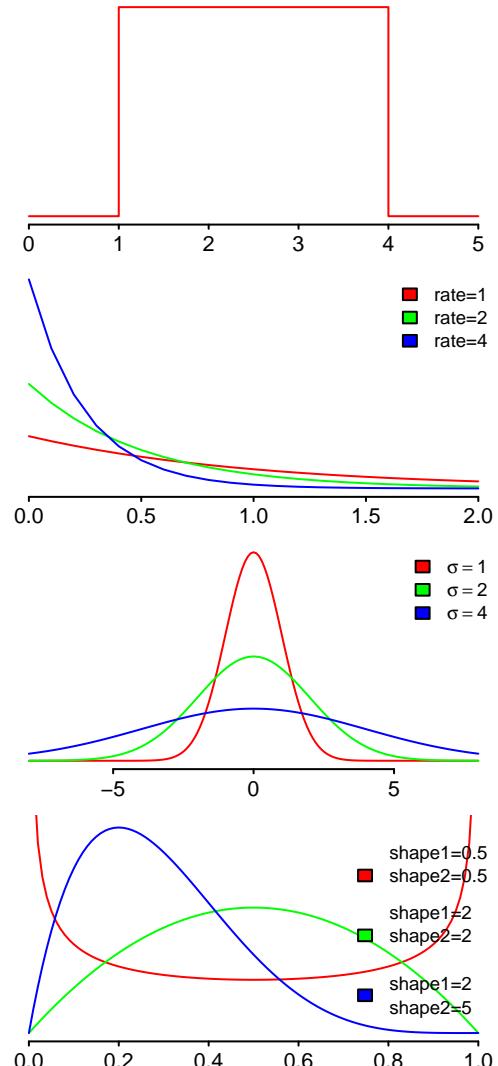


Figure 9.6: Commonly encountered continuous probability distributions (upper to lower: Uniform, Exponential, Normal, beta). [code](#)

- Cauchy distribution: this distribution is more famous for its unusual characteristics, e.g., having an undefined mean and variance (because of its very fat tail), than because of its uses. The density function for the average of two random variables each having a Cauchy density is a random variable with a Cauchy density; this self mapping is unique to the Cauchy distribution. One consequence is that if the error in a measurement has a Cauchy density, then the average of many measurements will not be more accurate than the individual measurements.

A chi-squared test is the traditional technique for testing whether data has a multinomial distribution, but has a higher false positive than an exact calculation. Use to a computer and the

### 9.2.1 Comparing probability distributions for equality

The question of interest is actually whether two or more samples are drawn from the same population, but the mathematics of sample comparison is framed in terms of comparing the measurable characteristics of probability distributions. These measurable characteristics include mean, standard deviation, skew and kurtoise (a probability distribution is defined in terms of a formula and parameters that get plugged into this formula; a formula is not a measurable).

Here we are interested in the distribution as a whole, not individual characteristics. The various comparison techniques are based on some measure of difference between the *shape* of the sample distributions.

As always visualization is a useful first step in obtaining information about whether two samples might be drawn from the same distribution. However, be warned that small datasets can produce visualizations showing little resemblance to the distribution from which they were drawn, as can be seen from Figure 9.7, where all the samples are drawn from the same Normal distribution.

A study by Veytsman and Akhmadeeva<sup>1403</sup> measured subject reading rate, in words per minute, for text printed using a Serif or Sans Serif font. Words per minute is a discrete distribution and subject performance is likely to be similar, i.e., there will be many duplicates. Figure 9.8 shows a density plot of the normalised data.

The following tests are based on comparing the edf (empirical distribution function) of the sample values.

- The Anderson-Darling test is based on the largest difference between the edf of the two distributions, it uses weights to ensure that the tails of the distribution have as much influence as other parts of the distribution; it is possible to use this test to compare more than two distributions. While the Kolmogorov-Smirnov test is often encountered it has been found to be less sensitive than the Anderson-Darling test<sup>1310</sup> because it primarily detects differences in the main body of the distribution, rather than over the complete range of values.

The `ad.test` function in the `kSamples` package implements the Anderson-Darling test for two or more samples.

The `ks.test` function, part of the base system, implements the Kolmogorov-Smirnov test; other implementation include the `ks.test` function in the `dgof` package whose interface is the same but includes support for discrete distributions.

Samples drawn from a continuous distribution are very unlikely to contain identical values and many implementations warn if a sample contains duplicate values.

- The Cramér-von Mises test is based on summing (the square of) differences between edfs, rather than using a single maximum value, and can be more powerful against a large class of alternative hypothesis.<sup>54</sup>

The `cvm.test` function in the `dgof` package implements the Cramér-von Mises test.

The choice of statistical test depends on whether differences over the range of values in the samples are of interest, whether tail values are uninteresting (perhaps because there are few measurements in the tail and so what is there is noisy) or the amount of difference between samples is the primary differentiator.

Support for comparing samples drawn from discrete distributions is provided by: the WRS package (on Github) implements a version of the Kolmogorov-Smirnov test (the ks function) that supports discrete data and also the bmpmul function which uses the Brunner-Munzel test; the ks.test function in the dgof package.

The following code performs various tests checking whether the two sample are likely to have been drawn from the same population (see [group-compare/tb104veytsman-dist.R](#)):

```
library("dgof")
library("kSamples")
library("WRS")

# From WRS
ks(serif$Standard_WPM, sansserif$Standard_WPM)
# In fact unscaled measurements give the same result, i.e., not different
ks(serif$WordsPerMinute, sansserif$WordsPerMinute)

dgof::ks.test(serif$Standard_WPM, ecdf(sansserif$Standard_WPM))

# From base system
ks.test(serif$Standard_WPM, sansserif$Standard_WPM)

# Only applicable to continuous distributions
ad.test(serif$Standard_WPM, sansserif$Standard_WPM)
```

In this case the hypothesis that the samples are drawn from different distributions is rejected.

Practical manual techniques often require that samples be drawn from a Normal distribution and this test commonly performed by other people.<sup>iii</sup>

The result of testing whether a small sample is drawn from a Normal distribution has a high degree of uncertainty. Figure 9.9 was obtained by testing samples of various sizes, all drawn from the same distribution (e.g., a call to rexp), using the shapiro.test function. The y-axis shows the probability of the Shapiro-Wilk test detecting ( $p\text{-value} < 0.05$ ) that the sample values are not drawn from a Normal distribution; for the case when the values are drawn from a Normal distribution (e.g., a call to rnorm) the y-axis gives the probability of this fact not being detected.

Note that many points may be needed to distinguish a difference when one exists. For instance, two samples of 150 points are needed to obtain a 95% confidence level, using ad.test, that the samples are drawn from different distributions, when one sample is drawn from an Exponential distribution and the other from a Normal distribution (550 points are needed when the samples are drawn from Normal and Uniform distributions); see [group-compare/ad-check.R](#).

There is no guarantee that the values in a sample will have a distribution that even closely resembles any of the known probability distributions.

A study by Berger, She, Czarnecki and Wąsowski<sup>128</sup> investigated the use of feature macros used in the configuration of software product lines. Figure 9.10 shows the number of conditionally compiled sections of source code that were dependent on a given number of feature macros.

The Cullen and Frey graph, produced by descdist, shows that the characteristics of neither sample are close to matching any of the common discrete distributions. A Kolmogorov-Smirnov test considers them to be sufficiently different that they are likely to have been drawn from different distributions (see [group-compare/cond-compile/2010-berger.R](#)).

Samples may appear to have a similar shape, but have different mean values. Technically, samples with different mean values (or standard deviations) are considered to be drawn from different distributions. There may be theoretical reasons for believing that samples have been generated by the same processes and normalizing mean values (or even variance) is of enabling the shape of the sample distributions to be compared.

A study by Zhu, Whitehead, Sadowski and Song<sup>1497</sup> counted the number of various kinds of statements in a corpus of C, C++ and Java programs (approximately 100 programs, around 10 million lines, for each language). Figure 9.11 shows the distribution of occurrence (expressed

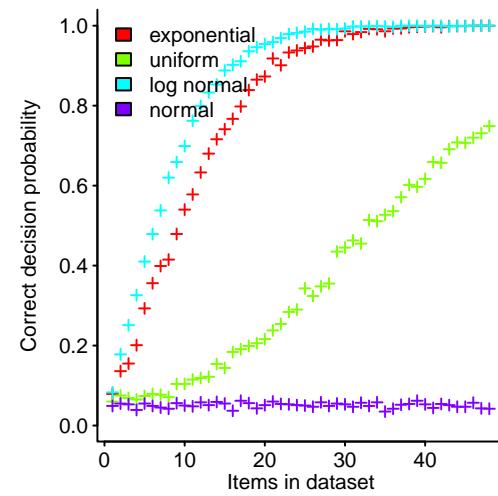


Figure 9.9: Probability, with 95% confidence, that shapiro.test correctly reports that samples drawn from various distributions are not drawn from a Normal distribution, and probability of an incorrect report when the sample is drawn from a Normal distribution. [code](#)

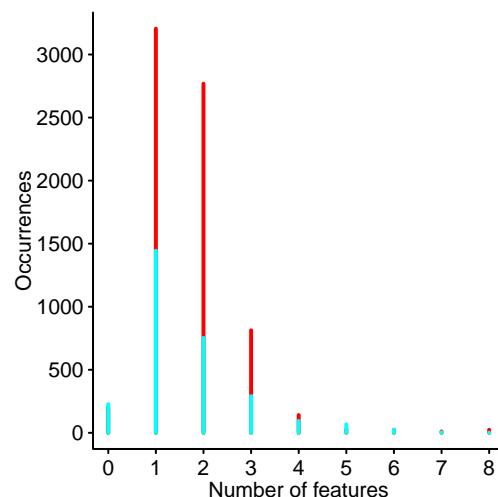


Figure 9.10: Number of conditionally compiled code sequences dependent on a given number of feature macros (red overwritten by blue: Linux, blue: FreeBSD). Data from Berger et al.<sup>128</sup> [code](#)

<sup>iii</sup> Readers of this book know about more powerful techniques that do not have this precondition.

as a density on the y-axis) of various statements (expressed as a percentage on the x-axis) over the programs measured; a different color for each language, figure out which is which before looking at the code.

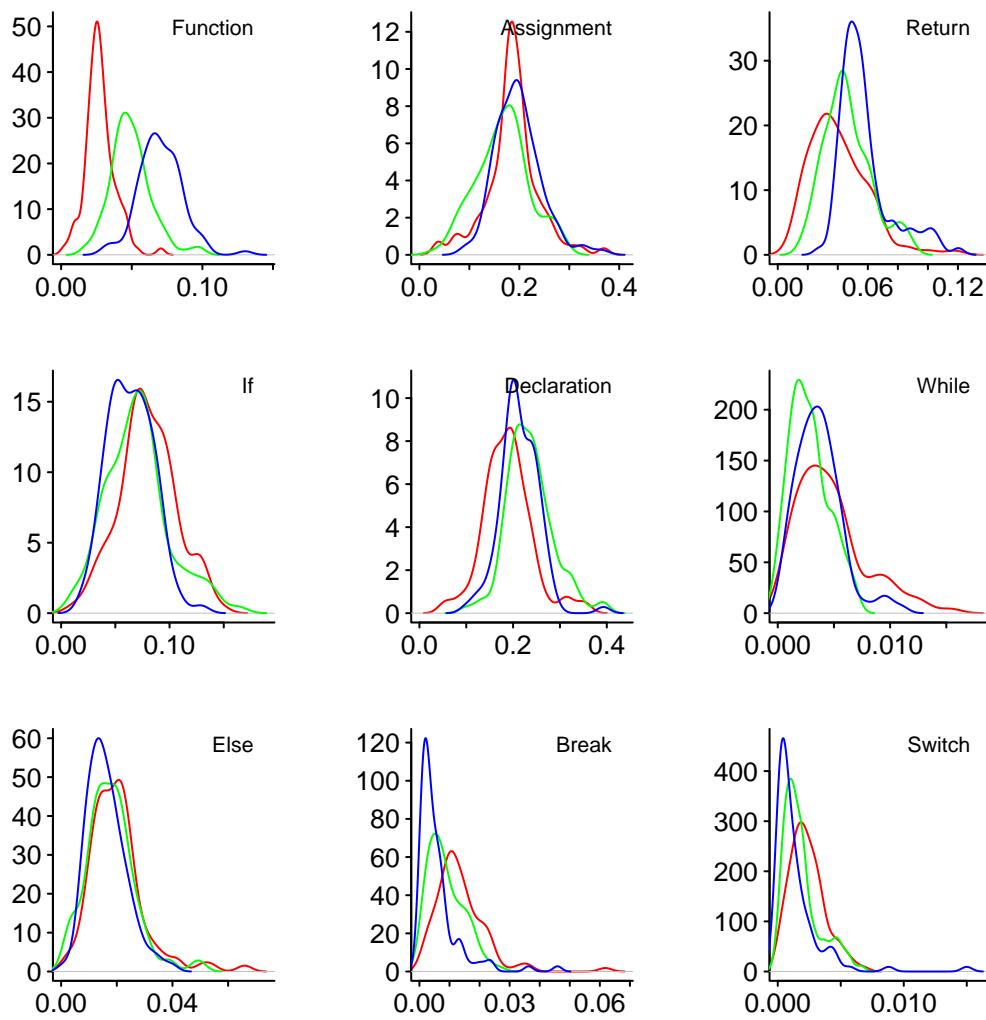


Figure 9.11: Percentage occurrence of statements for each of 100 or so C, C++ and Java programs, plotted as a density on the y-axis. Data from Zhu et al.<sup>1497</sup> code

Developers are pushed towards using particular statements by the characteristics of the application/language/algorithim. Differences in the probability of various kinds of statements being used, over a sample of programs written in various languages, is evidence that language has an impact on what code gets written (either because particular kinds of applications are written using a given language, particular algorithm selection is influenced by language, or the impact of differences in language semantics).

Might two or more of the languages be said to have the same distribution of if-statement and/or assignment-statement usage? The interactions between different statements makes the analysis non-trivial.

The takeaway from this section is that for small sample sizes distribution comparison produces unreliable answers and for large samples comparison may be complicated.

Comparison of particular characteristics of sample distributions, e.g., sample means, is covered in Chapter 13.

### 9.3 Fitting a probability distribution to a sample

Given a sample of values of a single attribute, which of the known, supported by R,<sup>255</sup> probability distributions is the best fit?

There is no universal best-test statistic for goodness-of-fit of a sample to a probability distribution. The performance of the available tests depends on the (unknown) distribution from which the sample was drawn.<sup>1303</sup>

The Normal distribution is often the default answer given, when people are asked about the distribution of a sample. There are several reasons for this, including: historically many of the techniques were designed to be performed by a human calculator were derived from theory that assumed normally distributed data (which often appeared to work reasonably well when the data only approximated a Normal distribution) and a misunderstanding of what the Central Limit theorem is about driving a belief that measurements of a complex process provides the mixing needed to produce a Normal distribution.

As always, knowledge of the processes driving the production of the measured values can be very useful. For instance, measurements of arrival times that are driven by a Poisson process will result in inter-arrival times that are exponentially distributed, values created via the multiplicative effect of many contributions may have a Lognormal distribution and a preferential attachment process often results in links or what they link following a power law.

If there is no theoretical justification for a particular distribution, limiting the selection process to those distributions having some degree of name recognition is likely make the final choice an easier sell to readers. For instance, the Delaporte distribution<sup>iv</sup> might happen to fit a particular sample slightly better than the Negative Binomial distribution, but its lack of name recognition will mean that extra effort will have to be invested on justifying its use.

A study by van der Meulen<sup>1389</sup> posted the  $3n + 1$  problem on a programming competition website: 95,497 solutions were submitted and van der Meulen kindly sent me a copy of these solutions (11,674 solutions were written in Pascal, the rest written in C). The  $3n + 1$  problem is to write a program that takes a list of integers and outputs the *length* of each value, where length is the number of iterations of the following algorithm:

```
for input integer ++pass:[n]++
  while (n != 1)
    n = (is_even(n) ? n/2 : n*3+1);
```

Which distribution is a good approximation to the number of lines of code contained in the programs submitted as answers to this problem?

The first step of visualizing the sample provides basic information about the shape of the distribution, e.g., decreasing/increasing, single/multiple peak, symmetric/skewed or appearing to be nothing but random noise.

One technique for narrowing down the list of possible distributions is to plot its Cullen and Frey graph. The `descdist` function in the `fitdistrplus` package plots this graph and returns some descriptive distribution characteristics of the values (mean, median, sd, skewness and kurtosis). Skew and kurtosis are not reliable estimators and `descdist` includes an option to create and test bootstrap samples.

The blue and yellow points in Figure 9.12 denote the sample and various bootstrap results for the  $3n + 1$  program lengths, assuming a continuous distribution (the average number of lines is large enough that the difference between discrete/continuous is likely to be very small).

```
library(fitdistrplus)

# Default is to check continuous distributions
# dummy=descdist(li, discrete=TRUE, boot=500)
dummy=descdist(li, boot=500)
```

The `fitdist` function<sup>v</sup> in the `fitdistrplus` package can be used to fit a distribution to the data, i.e., find values of the specified distribution's parameters, such as mean and variance, that minimise some measure of goodness-of-fit (the AIC of the fit is returned). The `gamlss` package supports a wider range of distributions (the help information for the `gamlss.family` function lists them) that can be used by `fitdist` to fit data.

Figure 9.13 shows fits for the Normal, Poisson, Lognormal and Negative binomial distributions.

```
library(fitdistrplus)

tp=fitdist(li, distr="pois")
```

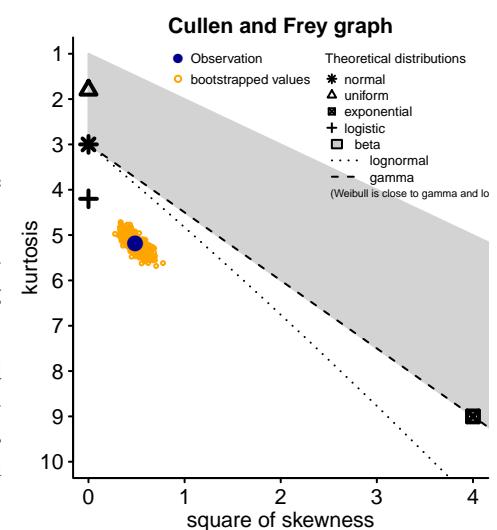


Figure 9.12: A Cullen and Frey graph for the  $3n + 1$  program length data. Data kindly provided by van der Meulen.<sup>1389</sup> code

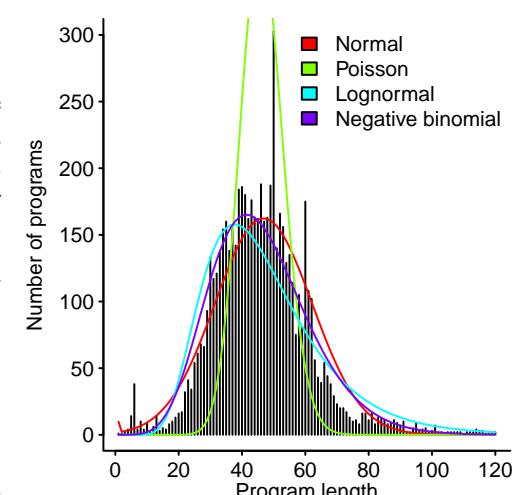


Figure 9.13: Number of  $3n+1$  programs containing a given number of lines and four distributions fitted to this data. Data kindly provided by van der Meulen.<sup>1389</sup> code

<sup>iv</sup> A compound distribution derived from a Poisson distribution whose mean has a shifted Gamma distribution.

<sup>v</sup> The MASS package contains the `fitdistr` function and the `gamlss` package contains the `fitDist` function both of which fit distributions to data.

```

tn=fitdist(li, distr="norm")
tln=fitdist(li, distr="lnorm")
tnb=fitdist(li, distr="nbinom")

# gofstat is a way of getting all the values used for plotting
theo_vals=gofstat(list(tn, tp, tln, tnb), chisqbreaks=1:120,
                   fitnames=c("Normal", "Poisson",
                             "Lognormal", "Negative binomial"))

plot_distrib=function(dist_num)
{
  lines(theo_vals$chisqbreaks, head(theo_vals$chisqtable[, 1+dist_num], -1),
         col=pal_col[dist_num])
}

plot(theo_vals$chisqbreaks, head(theo_vals$chisqtable[, 1], -1), type="h",
      xlab="Program length", ylab="Number of programs\n")
plot_distrib(1)
plot_distrib(2)
plot_distrib(3)
plot_distrib(4)

```

The large spike at 50 lines might be caused by a group of solutions all doing the same thing but with different statement orderings or multiple submissions derived from a common solution.

Based on minimizing AIC the Normal distribution is the best fit, with the Negative binomial distribution a close second. Should either distribution be chosen as the best fitting, or is it worthwhile attempting to fit other distributions? The answer depends on what the fitted distribution will be used for, e.g., making predictions or building models. Jumping to any conclusions based on one data-point (i.e., set of length measurements for one problem) is always problematic.

### 9.3.1 Zero-truncated and zero-inflated distributions

While zero is a common lower bound for measurement values, other lower bounds occur, e.g., the number of minutes to complete a task (the zero time tasks, that are never started, are not measured). Many potentially useful distributions describe variables whose values start at zero, e.g., the Poisson distribution.

It is possible to adjust the equations that describe zero-based distributions to have a non-zero lower bound. Rebasing a distribution to start at one (rather than zero) is the common case and after such an adjustment the distribution is said to be *zero-truncated*, e.g., *zero-truncated Poisson distribution*.

The `gamlss.tr` package contains functions that support the creation of zero-truncated (or truncation to the right or left of any value) distribution functions. The following call creates a set of functions relating to the zero-truncated type II Negative binomial distribution; the name of the created function is `NBIItr` and like other distribution functions in R there the associated density, distribution, quantile and random functions are obtained by prefixing the letters `d`, `p`, `q` and `r` respectively to `NBIItr`:

```

library(gamlss)
library(gamlss.tr)

gen.trun(par=0, family=NBII) # Bring various functions into existence

```

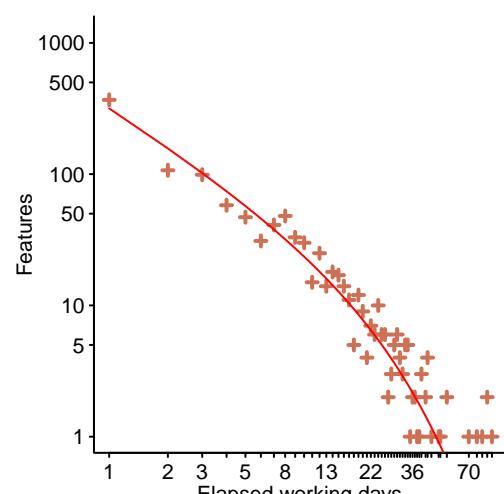
The 7Digital data<sup>1162</sup> (discussed in more detail in Section 5.4.5) contains information on 3,238 features implemented between April 2009 and July 2012; the information consists of three dates (Prioritised/Start Development/Done) from which a non-zero duration can be calculated.

The Cullen and Frey graph suggests a negative binomial distribution as a good fit (see ??).

The functions returned by `gen.trun` do not have a form that can be used in calls to the `fitdist` function. The `gamlss` function in the `gamlss.tr` package has a special form for handling these created functions, as shown in the following call (where `day_list` contains the list of values and `NBIItr` was created by an earlier call to `gen.trun`). The following code was used in the production of Figure 9.14:

v 0.10.0

Figure 9.14: A zero-truncated Negative Binomial distribution fitted to the number of features whose implementation took a given number of elapsed workdays; first 650 days used. Data kindly provided by 7digital.<sup>1162</sup> [code](#)



```

library(gamlss)
library(gamlss.tr)

g.NBIItr=gamlss(day.list ~ 1, family=NBIItr)

NBII.mu=exp(coef(g.NBIItr, "mu"))      # get mean coefficient
NBII.sigma=exp(coef(g.NBIItr, "sigma")) # standard deviation

plot(table(day.list), log="xy", type="p", col=point_col,
      xlab="Elapsed working days", ylab="Features\n")

lines(dNBIItr(1:93, mu=NBII.mu, sigma=NBII.sigma)*length(day.list),
      col="red")

```

One process generating values having a Negative binomial distribution is based on a mixture of Poisson distributions whose means have a Gamma distribution. It is possible to generate other distributions by combining a mixture of Poisson distributions, are any of these a better fit of the data? The Delaporte distribution sometimes fits slightly better and sometimes slightly worse (see Section 6.4.3 for details); the difference is not large enough to warrant switching from a relatively well-known distribution to one that is rarely covered in text books or supported in software; if data from other projects is best fitted by a Delaporte distribution then it may be worthwhile spending time analysing how this distribution might be a better model of project scheduling.

If the processes generating these values involve a mixture of Poisson distributions, then there unlikely to be a single subprocess responsible for a large percentage of the behavior, many subprocesses are involved.

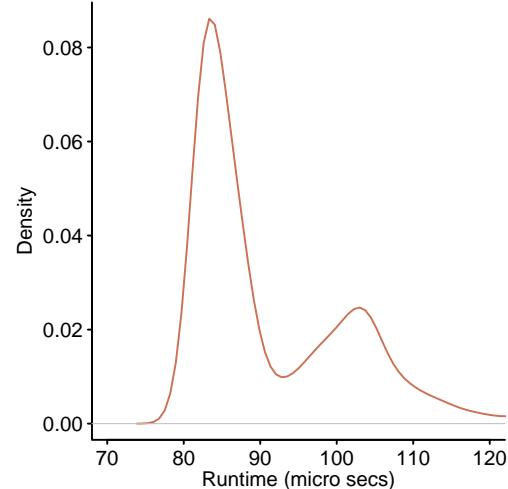
Sometimes count data measurements contain many more zero values than are expected from the discrete distribution that the generating process is believed to be following. Two of the processes that can generate extra zeroes include:

- a subset of the population is not involved in the process measured and so always has a zero value. This situation can be modeled by what is known as *zero-inflated model*, which combines a model of always zero vs. maybe non-zero values and a model of maybe non-zero values.

The `gamlss` package...

- the measurements involve two processes, one in which the values are zero or non-zero, and the other where values are always non-zero.

Just need some data... have emailed a few people... `probability/bolz_data_struct_racket.R` does not fit very well...



### 9.3.2 Mixtures of distributions

Sometimes sample values are generated by two or more distinct processes, resulting in measurements that appear to be drawn from two or more distinct distributions, e.g., a plot shows multiple peaks. A model built using a mixture, or weighted sum, of distributions is known as a *finite mixture model* or just a *mixture model*; a continuous mixture of distributions is known as a *compounded distribution* (the Negative Binomial distribution is a common compounded distribution).

The `mixtools` and `rebmix` packages contain functions for fitting samples drawn from two or more distributions, but they have to be from the same distribution family, e.g., multiple Normal distributions. The two packages differ in the structure of their API, e.g., one having many functions and the other one main function taking many arguments (neither would win a prize for user interface design).

A study by Hunold, Carpen-Amarie and Träf<sup>654</sup> investigated the impact of external factors on the performance of an MPI micro-benchmark. Figure 9.15 shows the runtime variation of two MPI calls, with each having two distinct peaks.

The lower Figure 9.15 shows two peaks with each appearing to be symmetrical and perhaps a mixture of two Normal distributions is a good fit. Figure 9.16 shows the two distributions fitted by a call to the `normalmixEM` function (in the `mixtools` package).

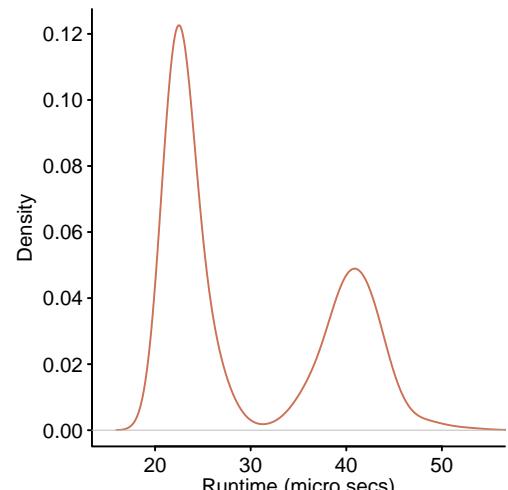


Figure 9.15: Density plot of MPI micro-benchmark runtime performance for calls to `MPI_Scan` with 10,000 Bytes (upper) and to `MPI_Allreduce` with 1,000 Bytes (lower). Data kindly supplied by Hunold.<sup>654</sup> code

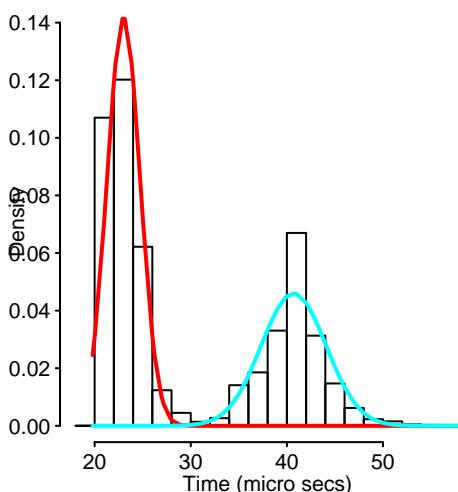


Figure 9.16: Mixture model fitted by the `normalmixEM` function to the performance data from calls to `MPI_Allreduce`. Data kindly supplied by Hunold.<sup>654</sup> [code](#)

```
library("mixtools")
scan_dist=normalmixEM(fig1_Allreduce$time)
plot(scan_dist, whichplots=2, main2="", col2=pal_col,
      xlab2="Time (micro secs)", ylab2="Density\n")
```

A call to `summary` returns the parameters of the fitted model. The first row (prefixed by `lambda`) is the fraction contributed by each distribution, followed by the mean, standard deviation and log likelihood (rather than AIC): [code](#)

```
number of iterations= 14
summary of normalmixEM object:
      comp 1   comp 2
lambda 0.611002 0.388998
mu     23.011364 40.703293
sigma   1.720527 3.378666
loglik at estimate: -28873.39
```

A plot of a sample drawn from a mixture of distributions does not always show visually distinct peaks.<sup>vi</sup>

A study by Kaltenbrunner, Gómez, Moghnieh, Meza, Blat and López<sup>726</sup> analysed the pattern of user activity of the Slashdot technical community news site. The black line in Figure 9.17 shows the density of the number of accesses to one article in each minute after first publication (a total of 1,567 accesses).

A possible explanation for the multiple upticks in number of accesses is the article being linked to from other web sites, driving a fresh batch of readers to Slashdot; the number of linking websites is not known, but this explanation sounds reasonable enough to build a mixture model containing a relatively large number of different distributions. Which mixture of distributions might best fit the access times of this Slashdot article? The Poisson distribution is often used to model arrival times and is the obvious first choice, but in this particular case turns out not to provide the best fit.

Figure 9.17 shows several Normal distributions fitted to data, on a log scale, using functions from the `rebmix` and `mixtools` packages. The algorithms used by packages do not guarantee to find the globally optimal solution and differences in the mix of distributions selected can occur because of differences during the search process.

```
library("rebmix")
slash_mod=REBMIX(Dataset=list(data.frame(users=log(slash$users))),
                  Preprocessing="histogram", cmax=5,
                  Variables="continuous", pdf="normal", K=7:45)

plot_REBMIX_dist=function(dist_num)
{
  y_vals=dnorm(x_vals, mean=as.numeric(slash_mod$Theta[[1]][2, dist_num]),
                sd=as.numeric(slash_mod$Theta[[1]][3, dist_num]))
  lines(x_vals, slash_mod$w[[1]][1, dist_num]*y_vals, col=pal_col[dist_num])
}
```

```
plot(work_den, main="", xlim=c(0, 10), ylim=c(0, 0.36),
      xlab="", ylab="Access density\n")
plot_REBMIX_dist(1)
plot_REBMIX_dist(2)
plot_REBMIX_dist(3)
plot_REBMIX_dist(4)
```

Fitting a Normal distribution to log scaled data means that the sample actually has a Lognormal distribution. Is the Lognormal distribution a good representation of the processes driving readers to access Slashdot articles? As always in model building the answer depends on what

<sup>vi</sup>If  $f_1$  and  $f_2$  are normal densities with means  $\mu_1$  and  $\mu_2$ , respectively, and both have the same variance  $\sigma^2$ , then the mixture density  $f = 0.5f_1 + 0.5f_2$  will have a single peak if, and only if,  $abs(\mu_2 - \mu_1) \leq 2\sigma$ .

the model will be used for. If the purpose is to make predictions, then the accuracy of prediction is of more interest than any underlying assumptions. If the purpose is to understand what is going on, then a theory that contains processes generating Lognormal distributed behavior is needed.

It can take a lot of analysis over many years to settle on the distribution, or combinations of distributions, that best describes the measured properties of some system. The study of file-system characteristics<sup>14</sup> is an example of how researchers' ideas and models changed over time,<sup>360,668,962</sup> becoming more sophisticated as more data became available, from various platforms,<sup>318</sup> and more analysis was carried out.

### 9.3.3 Heavy/Fat tails

*Heavy tailed* is the term used to describe distributions where the majority of measured events occur a long way from the mean value (*fat tails* and *long tail* are also used).<sup>vii</sup> When the 80/20 rule applies the distribution is heavy tailed and the frequency with which this rule is encountered suggests that such data is not rare.

The `poweRlaw` package supports operations involving a variety of heavy tailed distributions, including power laws.

The Pareto distribution is the mathematical name of a particular instance of a heavy tailed distribution (sometimes going by the name *power law* in popular culture); Zipf's law is a particular instance of this distribution.

The mean of a heavy tailed distribution may not exist (because it is infinite). Any finite dataset has a finite mean, and if a sample is drawn from a heavy tailed distribution, its mean value will jump around erratically.

It is more difficult to narrow down which distribution might best fit a sample drawn from a heavy tailed distribution (because several fit equally well), compared to one without a heavy tail, because measurements are spread out rather than clumped around a central location.

Figure 9.18 is from a survey<sup>668</sup> of file sizes and shows that a small percentage of files account for most of the disk space occupied (the vertical line meets the bytes line where 89.9% of disk space has yet to be consumed and the files line where 12.5% of files still remain to be accounted for). Another way of describing the situation is to say that there is a mass/count disparity, i.e., a few files occupy most of the space.

Care needs to be taken to separate out concepts that are popularly associated with power laws, e.g., *scale invariant*, that are a property of the distribution, not the generating process. The process generating data that fits a power law can be remarkably random, e.g., the length of words in text produced by monkeys typing.<sup>283</sup>

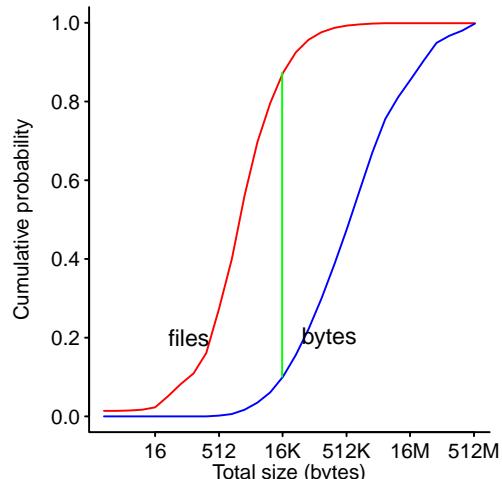


Figure 9.18: Cumulative probability distribution of files size (red) and of number of bytes occupied in a file system (blue). Data from Irlam.<sup>668</sup> code

## 9.4 Markov chains

A finite state machine (FSM) is a machine represented by a set of distinct states connected by edges denoting the possible transitions that can occur when a given event occurs, such as when a particular character is input (FSMs are deterministic).

A Markov chain (MC) is also a machine represented by a set of distinct states connected by edges, but the possible transition is chosen at random based on the transition probability of each edge (the transition probabilities out of any state, that is not an absorbing state, add to one); the next state only depends on the current state, i.e., the system is memoryless.

A Markov chain is a *discrete-time Markov chain* (DTMC) if the transition between states occurs at fixed time intervals; if the time interval between state transitions is not fixed, the Markov chain is a *continuous-time Markov chain* (CTMC) (the memoryless requirement means that transition times must have an exponential distribution). If the transition time depends on how long the system has been in the current state, it is a semi-Markov process (SMP).

Multi-state models, the `msm` package... the `markovchain` package...?

<sup>vii</sup> The term *sub-exponential* is sometimes used to describe tails that decay slower than exponential and *super-exponential* for tails that decay faster than exponential.

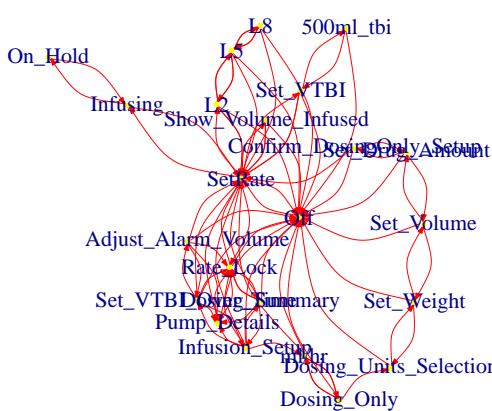


Figure 9.19: Graph of available state transitions for Alaris volumetric infusion pump (the button presses that cause transitions between states are not shown). Data kindly supplied by Oladimeji.<sup>1030</sup> code

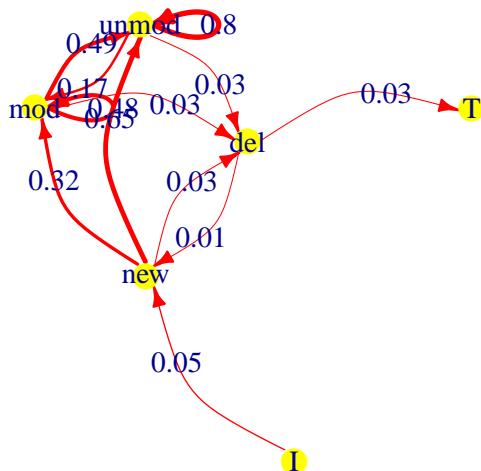


Figure 9.20: Discrete-time Markov chain for created/modified/deleted status of Linux kernel files at each major release from versions 2.6.0 to 2.6.39. Data from Tarasov.<sup>1339</sup> code

Finite state machines provide a useful abstraction for modelling user interfaces. A study by Oladimeji<sup>1030</sup> investigated the user interface of the Alaris volumetric infusion pump (a medical device used for controlled automatic delivery of fluid medication, or blood transfusion, to patients); the user interface includes 14 buttons and an LCD display. Figure 9.19 shows the available transitions between states.

Representing the FSM as a control flow graph, functions in the `igraph` package can be used to answer questions such as: the maximum number of button presses needed to get to any state (12; the `path.length.hist` function returns a count of all possible path lengths) and the average number of presses to transition between any two states (4; using the `average.path.length` function).

If the behavior of a system (that can be represented using a FSM) is monitored, the probability of occurrence of every transition between states can be calculated. If the behavior represents typical user interaction with the system, then the probabilities can be used to create a Markov chain for this typical behavior.

A study by Tarasov, Mudrankit, Buik, Shilane, Kuennen and Zadok<sup>1339</sup> used data on the lifetime of source files in various systems, such as the Linux kernel, to generate realistic filesystem contents (for deduplication analysis). Figure 9.20 shows a Markov chain for the source files in the Linux kernel (from being Initialised to new, through modified/unmodified to deleted and reaching the Terminal state). The measurement snapshot occurred at each of the 40 releases between versions 2.6.0 and 2.6.39, with an average of 23k files per snapshot; the time between releases is roughly constant, so this might be considered a discrete-time Markov chain.

```
library("igraph")

atc=read.csv(paste0(ESEUR_dir, "probability/atc12-gra.csv.xz"), as.is=TRUE)
atc_gra=graph.data.frame(atc, directed=TRUE)

V(atc_gra)$frame.color=NA
V(atc_gra)$size=12 ; V(atc_gra)$color="yellow"
E(atc_gra)$arrow.size=0.5 ; E(atc_gra)$color="red"
E(atc_gra)$weight=E(atc_gra)$linux
E(atc_gra)$label=E(atc_gra)$weight/100

# layout.lgl outperforms the default layout for this graph
plot(atc_gra, edge.width=0.3*sqrt(E(atc_gra)$weight),
     edge.curved=TRUE, layout=layout.lgl)
```

The `graph.data.frame` function assumes there is a link between the row values in two columns (from and to vertices) and builds a graph using this information. The `V` and `E` functions access the vertices and edges of the graph and various attributes can be set and may be subsequently used by `plot`.

The algorithm used by `plot` to layout a graph makes use of randomization, which means that the layout returned by every call will be different.

### 9.4.1 A Markov chain example

A study by Perugupalli<sup>531, 1080</sup> investigated the reliability of gcc based on the reliability of its major subsystems. Information on the probability of a subsystem experiencing a failure was calculated using the regression suite for gcc version 3.3.3 (which contains tests for 110 faults present in gcc version 3.2.3, out of 2,126 tests, of which 55 were traced back to the source code of a single subsystem; the others faults involved multiple subsystems). The researchers did not attempt to analyse failures involving more than one subsystem and assumed that subsystems fail independently of each other.

Subsystems were identified by instrumenting gcc to count the number of calls between pairs of functions performed while executing the regression suite (this is not Markov chain-like behavior because the called functions return, which is not transition-like behavior). The 1,759 traced functions were manually assigned to one of 13 internal subsystems (e.g., parsing, tree optimization and register allocation),

The reliability of gcc version 3.2.3 might be estimated using:

$$R = 1 - \frac{F_c}{T_c} = 1 - \frac{110}{2126} = 0.948$$

where  $F_c$  is the number of source files that it did not correctly compile and  $T_c$  is the total number of files compiled.<sup>viii</sup>

This approach has the advantage of being simple to calculate, but it does not provide any information on the impact of individual subsystems on overall reliability, for instance, what is the sensitivity of overall system reliability to behavioral changes to one subsystem?

The probability of reaching subsystem  $n$  from subsystem 1 after  $k$  transitions is  $Q^k$  (where  $Q$  is the matrix of transition probabilities). Summing over all transitions (using an infinite upper bound for the total number of transitions simplifies the mathematics) we get:<sup>1364</sup>

$$S = \sum_{k=0}^{\infty} Q^k = (I - Q)^{-1}$$

where:  $I$  is the identity matrix. The expression  $(I - Q)^{-1}$  is easily calculated (i.e., inverting the result of a matrix subtraction).

The matrix  $S$  is known as the *fundamental matrix* and can be used to calculate a variety of properties of systems modeled by the Markov chain.

The composite and hierarchical methods are two techniques for combining information on subsystem usage (i.e., subsystem transition probabilities and subsystem reliability calculated using the above formula) to obtain a value for the reliability of a complete system:

- composite method:<sup>249</sup> calculates the probabilities a successful transition between each subsystem by multiplying the transition probabilities of each subsystem by the probability of the subsystem executing successfully. These individual successful transition probabilities are used to calculate the successful transition probability from the initial subsystem to the final subsystem (i.e., the systems fundamental matrix). The estimated reliability calculated for gcc is 0.9972,<sup>ix</sup> (see [reliability/gcc-reliability.R](#)).
- hierarchical method: if  $R_i$  is the reliability of a subsystem, the probability of all executions of that subsystem being successful is  $R_i^{N_i}$ , where  $N_i$  is the number of transitions to subsystem  $i$  during one execution of the system. Assuming that subsystems fail independently, the expected value of the system reliability is:

$$R = E \left[ \prod_{i=1}^n R_i^{N_i} \right]$$

Assuming subsystems are highly reliable and the variance in the number of subsystem transitions is very small, the first order Taylor approximation can be used:

$$R \simeq \prod_{i=1}^n R_i^{V_i}$$

where  $V_i = E[N_i]$  is the expected number of times a transition occurs to subsystem  $i$  during a single execution of the complete system;  $V_i$  is obtained by solving:

$$V_i = q_i + \sum_{j=1}^n V_j p_{ji}$$

where  $q_i$  is the probability that execution starts with subsystem  $i$  and the  $p_{ji}$  are obtained from the subsystem transition probability matrix (see [reliability/gcc-reliability.R](#)).

There is a plentiful supply of books discussing the use of Markov chains to solve problems. The kinds of problems attacked using a Markov chain approach don't occur very often in this book.

## 9.5 Social network analysis

The popularity of web based social networks has made the mathematics of social network analysis a fashionable research topic. Unfortunately many of the results involve little more

---

<sup>viii</sup> The calculated reliability is very low because it is based on compiling a test suite of short code samples designed to reveal faults.

<sup>ix</sup> If the 55 fault count used in this analysis is plugged into the formula used above, the reliability estimate is 0.974.

than claiming to have found a power law, with only pretty pictures and blustering hand waving to show.

An example of the kind of descriptive statistics encountered in a social network analysis appears in Section 6.4.3 and the section on regression modeling discusses building models based on network data.

Social networks are represented as graphs and the `igraph` package supports reading the common graph data representation formats, along with a wide range of operations and analysis on graphs.

Both FreeBSD and OpenBSD were forked from a common base and not only continue to share common code but faults fixed in one are often applied, some time later, to the other. A study by Canfora, Cerulo, Cimitile and Di Penta<sup>217</sup> analysed the developer's mailing lists for these systems to obtain information on what they called *Cross-System-Bug-Fixings*; the data contains information on 861 unique developers sending email and 1,062 unique developers receiving email. Figure 9.21 was produced using code very similar to that used earlier for the Markov chains.

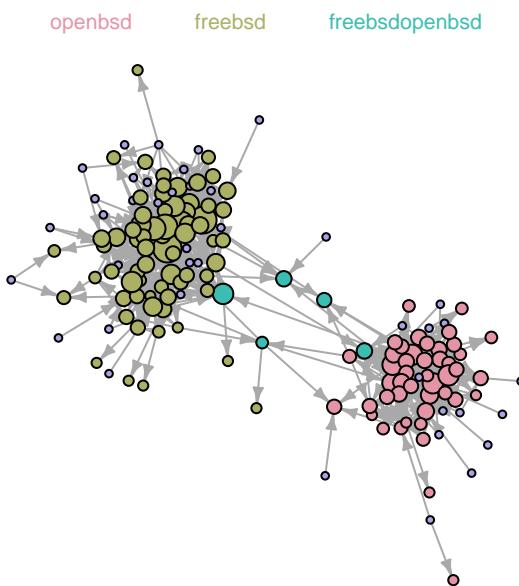


Figure 9.21: Directed graph of emails between FreeBSD and OpenBSD developers, plus a few people involved in both discussions, with developers who sent/received less than four emails removed. Data from Canfora et al.<sup>217</sup> code

Many real world linked collections contain subgroups (e.g., clusters of developers or related code modules) and there are a variety of algorithms for detecting these subgroups. Care needs to be exercised in interpreting the clusters returned by these algorithms as there may be many distinct high-scoring solutions and a clear global maximum may not exist.<sup>515</sup>

## 9.6 Combinatorics

The analysis of some systems makes it necessary to consider combinations of various items and there is a need to enumerate all possible sequences, to calculate the total number of different sequences of items that could occur or other related questions. The mathematics used to solve this kind of problem is known as *combinatorics*.

A few of the functions frequently used in combinatorial problem solving are included in R's base system, including:

- `choose` takes two arguments,  $n$  and  $k$  and returns the value  $\frac{n!}{k!(n-k)!}$ , often written as  $\binom{n}{k}$ ; the number of ways of selecting  $k$  items from  $n$  items,
- `combn` takes two arguments,  $x$  and  $k$  and returns an array containing all combinations of the elements of  $x$  taken  $k$  at a time.

When an item is drawn, with replacement, from a pool of items the probability of drawing the same item again is unchanged, when drawn without replacement the probability will decrease by the appropriate amount. An item is distinct if it is treated as being different from all other

items in the pool (even when drawing with replacement), e.g., there are four items in the pool `x=c("a", "a", "b", "c")`, but only three of them are distinct.

Table 9.1 show how the `iterpc` function in the `iterpc` package can be used to generate sequences based on the distinctness of the items and whether they are drawn with replacement or not.

Distinct			
	True	False	False
True	<code>I=iterpc(5, 2, replace=TRUE)</code>		<code>x=c("a", "a", "b", "c")</code>
			<code>I=iterpc(table(x), 2, replace=TRUE)</code>
Replacement			
False	<code>I=iterpc(5, 2)</code>		<code>x=c("a", "a", "b", "c")</code>
			<code>I=iterpc(table(x), 2)</code>

Table 9.1: Example `iterpc` calls generating particular kinds of sequences of length two (by passing the value returned to `getall`, e.g., `getall(I)`).

The treatment of item ordering is another factor when considering all possible permutations; is the ordering of items significant or not, e.g., are the sequences `a,b` and `b,a` treated as different or equivalent? When the ordering of items is significant calls to `iterpc` need to set the optional argument `ordered` to `TRUE`, e.g., `I=iterpc(5, 2, ordered=TRUE)`.

### 9.6.1 A combinatorial example

A study by Jones<sup>703</sup> investigated developer preferences for ordering the members of C `struct` types. The hypothesis was that members having the same type are grouped together within the same `struct` type.

There are enough instances of `struct` types containing between three and eight members in the sample to be analysed with a reasonable level of confidence.<sup>x</sup>

If a `struct` contains  $n$  members, then the number of possible member sequences is  $n!$ . However, we are only interested in member types and don't care about permutations of members having the same type. The number of different member type sequences is  $\frac{n!}{n_1!n_2!\dots}$  where  $n = n_1 + n_2 + \dots$  and  $n_1, n_2$ , etc are the number of members having a given unique type.

Taking the example of a `struct` containing four members, two of type `x` and two of type `y` the possible sequences of member types within a `struct` type are:

`xxyy` `xyxy` `yxxx` `xyyx` `xyyx` `yyxx`

and if two members are of type `x`, one of type `y` and one of type `z` the possible member type sequences are:

`xxyz` `xxzy` `xyxz` `xyzx` `xzxy` `yxxx` `yxzx` `yzxx` `zxyx` `zxxx` `zyxx`

In the first case members are grouped together in  $\frac{1}{3}$  of cases and in the second in  $\frac{1}{2}$  of cases.

If there are  $t$  different types, there are  $t!$  possible unique sequences of types. If the ordering of `struct` members is random, the probability of encountering a definition in which all members having the same type are grouped together is:

$$\frac{t!}{\frac{n!}{n_1!n_2!\dots n_t!}}$$

For the two examples above the probabilities of having member ordering where identical types are grouped together are:  $\frac{2!}{4!}$  and  $\frac{3!}{2!2!1!1!}$  (which we already knew from writing out all possible sequences).

When a `struct` contains four members, as in the above examples, it is not possible to distinguish between a developer intentionally choosing an order and random selection. Repeating the calculation for `structs` types containing five members shows that the probability of random selection of member order producing the same member types being grouped together is surprisingly high (see the fifth column in Table 9.2).

<sup>x</sup> The number of `struct` types containing a given number of members decreases approximately logarithmically with increasing number of members,<sup>699</sup> i.e., most member sequences are relatively short.

Total members	Type sequence	structs seen	Grouped occurrences	Random probability	Occurrence probability
4	1 1 2	239	185	0.50	$2.83 \times 10^{-18}$
4	1 3	185	146	0.50	$4.75 \times 10^{-16}$
4	2 2	98	61	0.33	$4.58 \times 10^{-9}$
5	1 1 1 2	57	50	0.40	$1.03 \times 10^{-13}$
5	1 1 3	94	61	0.30	$3.13 \times 10^{-12}$
5	1 2 2	86	49	0.20	$5.18 \times 10^{-14}$

Table 9.2: Various forms of struct types containing a given number of members, one possible type grouping, number of actual struct types measured, number having grouping, probability that one type will contain this grouping and probability that the number grouped out of total seen will be so grouped. Data from Jones.<sup>703</sup> code

Table 9.2 shows source code measurement counts and calculated probabilities for struct types containing four and five members: the column *Total members* lists the number of members in the type, *Type sequence* is a possible grouping of member types for the given number of members, *structs seen* is the number of measured structs containing the given number of members/types, *Grouped occurrences* is the number of measured structs having the grouping listed in the first column, *Random probability* is the probability that this grouping will occur randomly in one struct declaration containing that number of members and types, *Occurrence probability* is the probability that *Grouped occurrences* out of *structs seen* will occur when the probability of a single instance occurring is *Random probability*.

These calculations show that it is not possible to confidently distinguish between random and intentional ordering for individual struct types. However, programs contain many such types and if we label each one "Yes" or "No", depending on whether their member types are grouped or not, this list of Yes/No labels have a binomial distribution and the probability of the given number of Yes/No labels occurring through chance can be calculated.

Taking the example of a struct containing four members, two of type x and two of type y, the sample contains 98 such types with 61 of them having grouped member types (see columns 3 and 4 of Table 9.2). The probability of this occurring, when the random occurrence probability is  $\frac{1}{3}$ , is calculated using `pbinom(61-1, 98, 1/3, lower.tail=FALSE)`, whose value is  $4.58272e-09$  (the `lower.tail=FALSE` option is used because we are interested in the probability of seeing 60 or more occurrences).

- + Figure 9.22 shows the measured percentage of struct types whose members are grouped by type (red pluses) and the percentage that would occur with random ordering (blue line). The green line is the 99.9% probability bound for the likelihood that 100 structs, all sharing the same member types, will all have their members grouped by type when member ordering is chosen at random. The distance of the red crosses from the 99.9% bound strongly shows that grouping of members by type has not been driven by random selection.

## 9.6.2 Generating functions

Generating functions are discussed here purely to inform readers about a powerful technique that is significantly different from the traditional approach to solving probability problems using factorials; this technique is capable of solving problems that appear to be otherwise intractable. If you cannot derive an expression specifying how many possibilities can occur in some situation, then a search for the appropriate generating function may provide an answer.

Generating functions are starting to be covered in texts on probability; some mathematical sophistication is required.

A generating function is a polynomial  $a_0x^0 + a_1x^1 + \dots + a_nx^n$  where the coefficients  $a_n$  encode information about the quantity of interest.

The following is a simple example that could just as easily be calculated using factorials, but illustrates the idea. How many ways can five items be selected if A can be selected 0 or 1 times, B can be selected 0, 1 or 2 times and C can be selected 0, 1, 2, 3 or 4 times? The generating function is (see the suggested reading for why this works):

$$(1+x)(1+x+x^2)(1+x+x^2+x^3+x^4) = x^7 + 3x^6 + 5x^5 + 6x^4 + 6x^3 + 5x^2 + 3x + 1$$

the coefficient of  $x^5$  is 5, so five different items orderings are possible.

A more complicated example is when items have a particular value and sequences that sum to a specific total are required. If A is worth 1, B is worth 3 and C is worth 5, the generating function is:

$$(1+x+x^2+\dots)(1+x^3+x^6+\dots)(1+x^5+x^{10}+\dots) = 7x^{11} + 7x^{10} + 6x^9 + 5x^8 + 4x^7 + 4x^6 + \\ 3x^5 + 2x^4 + 2x^3 + x^2 + x + 1$$

the coefficient of  $x^{10}$  is 7, so there are seven different ways of selecting items that sum to ten.

The `polynem` package...



# Chapter 10

## Statistics for software engineering

### 10.1 Introduction

The output of statistical analysis should be treated as a guide and not a mandate.

Correlation does not imply causation, a common mantra that is worth repeating.

Traditionally, statistical techniques have had to be practical to perform manually. This has resulted in general problems being split into a profusion of specific subproblems and the creations of techniques tailored to do a good job of handling each case. Doing statistics involved mapping the sample characteristics to a particular subproblem and then applying the corresponding technique. Using a computer makes it practical to apply general solution techniques and there are a few more powerful and robust statistical techniques available.<sup>391</sup> many users of statistical techniques have simply switched from manual to computer based calculation of the familiar historical techniques, without appreciating the original design rational for these techniques. Many of the statistical techniques appearing in this book are impractical to apply manually, e.g., the bootstrap, a computer is required.

The developer input to the data analysis process is to apply domain knowledge to suggest possible patterns of behavior to search for and to provide one or more interpretations of any other patterns that might be uncovered.

This chapter attempts to illustrate the use of techniques likely to be of use in analysing data that is commonly encountered in a software engineering context. It starts by discussing some general ideas that are used in statistical analysis.

Many books using statistical techniques invest a lot of effort massaging data into a form that permits the use of techniques that assume the data has a Normal distribution, a.k.a. Gaussian distribution. The reasons for this are historical (assuming Normality made the analysis tractable in the days before computers) and data in the Social sciences (early adopters of statistical techniques and a huge market for statistical books) appearing to be drawn from a Normal distribution (despite the claims made, data in this field often does not have a Normal distribution<sup>939</sup>). It could be said that nobody ever got fired for assuming a Normal distribution.

Software engineering measurements often involve values that do not have a Normal distribution; the Exponential and Poisson distributions are relatively common; measurements best described using a Normal distribution do occur, but they do not have the dominant market share encountered in other, non-software related, domains (e.g., the social sciences).

The input to any statistical analysis is a sample and some expectations of behavior; the expectations may be explicit (e.g., measurements are independent of each other) or implicit assumptions (e.g., the choice of a statistical technique that only produces reliable results for samples drawn from a population having certain characteristics).

The terms *parametric test* and *nonparametric test* are sometimes used to describe statistical tests; a parametric test assumes that the sample has some known distribution (i.e., an expression containing configuration parameters such as mean and standard deviation), while a nonparametric test makes no assumptions about the distribution of the sample (it is sometimes said to be *distribution free test*).

When a sample has the required distribution, parametric tests have greater power for the same number of measurements relative to a corresponding nonparametric test.

Historically, the reliability of the results produced by statistical techniques has depended on mathematical analysis; analytic solutions are only known for a relatively small number of cases. Once computers became available they provided another tool that could be used to evaluate the performance of techniques on data having a wide variety of characteristics, e.g., how poorly some parametric techniques perform when a sample does not have the required probability distribution.

Ideally the person doing the analysis has a good understanding of the processes that generated the sample being analysed, this understanding can be used to create a model against which the measured values can be compared. In practice the understanding of the processes involved is often poor or non-existent and in this case the available sample is used to guide the analysis, which is almost as dangerous as not using it.

Having found a pattern that matches to some desired level of certainty, the next question is the size of the effect (e.g., mountain or molehill; the statistical term is *effect size*). If the effect size is large enough to be of interest, the next step is to obtain some assurance that it really exists and is not the result of some random combination of events. Confidence intervals and *p-value* are two of the quantities used to compare the results obtained against random behavior.

The ability to detect patterns that might be present in a sample depends on the quality and quantity of the measurement data:

- quality: noise in the measurement process and errors in post measurement processing (e.g., incorrect conversion of file formats or inaccurate calculations of values derives from the raw data) are some of the problems that affect the quality of the sample data,
- quantity: the number of measurements impacts the power and significance of statistical tests and the error bounds of statistical algorithms.

Some statistical techniques divide variables into two classes, those that explain and those that respond. As their name suggests *explanatory variables* are used to explain something (the term *predictor variables* is used when the variables are used to make predictions, *control variables* is sometimes used in an experimental setting and *independent, stimulus, factor* and are also used). The *response variable* (also known as a *dependent variable*) is the variable (usually there is only one) whose behavior is explained or predicted using explanatory variable(s).

### 10.1.1 Statistical inference

The most commonly used statistical inference technique is based on *frequentist* methods, i.e., how often events occur and long-run averages. All techniques have problems associates with their use and frequentist being the most widely used has the greatest number of detractors; a common problem is misuse of the concept of p-value, a problem likely to befall any widely used technique simply because of the varying skills of the people using it. The p-value is the fall-guy of the frequentist approach and the issues surrounding this quantity are discussed below.

The frequentist approach is the technique predominantly used in this book because it is commonly used in statistical books and articles, it is used by most of the R packages and readers are likely to encounter it when interacting with other people involved in analysing and using data.

Another inference technique is *Bayesian statistics*. A major issue with the Bayesian approach is that it requires an estimate for the probability of an event occurring, i.e., a reasonable value for the probability of the event occurring estimated prior to any measurements being made (the measurements get factored in later); selecting a suitable prior opens the door to the bias of opinion and policy guidelines,<sup>164</sup> e.g., a Bayesian approach to deciding whether the accused is guilty runs into the problem that many legal systems assume that people are innocent until proven guilty (i.e., the prior is zero), which steps through the calculation to always producing a non-guilty answer.

*Maximum likelihood estimation*, MLE, is a technique that finds the set of parameters for a model that makes the observed data most likely to have occurred. It has been said that academics like maximum likelihood estimation because it is a way of showing off their calculus within the statistics syllabus.

A study by Furia<sup>464</sup> reanalysed several software engineering datasets using Bayesian techniques...

## 10.2 Samples and populations

It is not always possible to measure every member of a population and the subset of a population measured is known as a *sample* (see Figure 10.1).<sup>i</sup> Depending on the question being asked, a set of measurements may be a population or a sample. For instance, measurements of one particular program yields the parameters of a population when the questions being asked concern just that one program, but they become the statistics of a sample when generalizing the findings to questions about other programs (including future versions of the one measured).

A sample is selected as a proxy of the entire population. There are a variety of sampling techniques, including:

- a *prospective* study collects data as events unfold. Figure 10.2 shows the date of introduction of a cpu against its commercial lifetime, in years.<sup>689</sup> Processors that ceased production in 2000 or 2010 would appear along one of the two colored lines,<sup>ii</sup>
- a *retrospective* study collects data after events have taken place,
- a *convenience sample*, as its name implies, makes do with what is available,
- *snowball sampling*, or *chain sampling* starts with an initial list of subjects who are asked to propose other subjects whom to them, with the process iterating until the number of new subjects falls below some threshold,
- stratified sampling divides the population into what are known as *strata*, with the strata chosen so that similar cases cluster tend to within each one; each of these strata are then sampled (using say random sampling) to produce the final sample (which is a set of distinct stratum, see Figure 10.3),
- sequential sampling is covered in Chapter 13,
- interval sampling divides the measurement interval into a series of fixed points and samples at just these points. The width of the sampling intervals puts a lower bound on the behavior that can be resolved. An experimental study<sup>iii</sup> by Kistowski et al<sup>1383</sup> measured power consumption, using programs from SPEC's Server Efficiency Rating Tool, at load level increments of 2% (crosses) and 10% (lines); see Figure 10.4. A cost/benefit analysis would compare the greater accuracy obtained using finer measurement intervals against the likelihood of sudden jumps in the value of the response that could have a noticeable impact on the results.

Occasionally the subjects of interest are not present in the sample. For instance, the damage experienced by aircraft returning from combat, during the second world war, was analysed with a view to improving aircraft survival rate. One of the statisticians involved pointed out that important subjects were missing from the sample,<sup>891</sup> aircraft that had not returned. The return of a damaged aircraft provided evidence that the damaged areas were not critical to survival; it was those areas not damaged in returning aircraft that were likely to be critical to survival.

Guy<sup>568</sup> proposes a *strong law of small numbers*: "There aren't enough small numbers to meet the many demands made of them."; he lists 35 examples of numeric patterns found in samples calculated using small integer values that disappear when larger integer values are used (i.e., the sample size is increased).

<sup>i</sup> The term *statistic* applies to values calculated from a sample, while the term *parameter* applies to values calculated from a population.

<sup>ii</sup> Email discussion with the author confirmed that the data had not been updated since 2010.

<sup>iii</sup> The study was experimental because it did not meet all the requirements for an official SERT run.

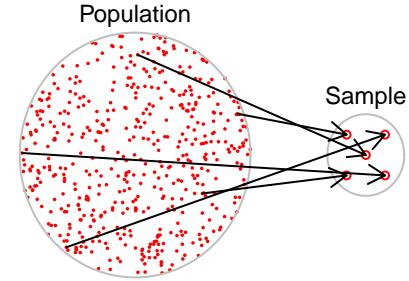


Figure 10.1: Example of a sample drawn from a population. [code](#)

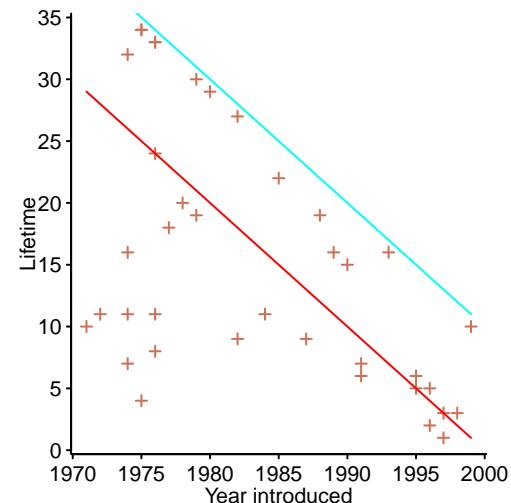


Figure 10.2: Date of introduction of a cpu against its commercial lifetime. Data from Culver.<sup>689</sup> [code](#)

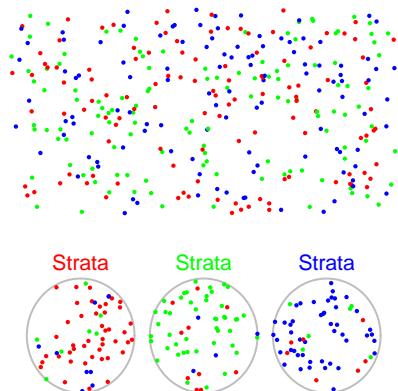
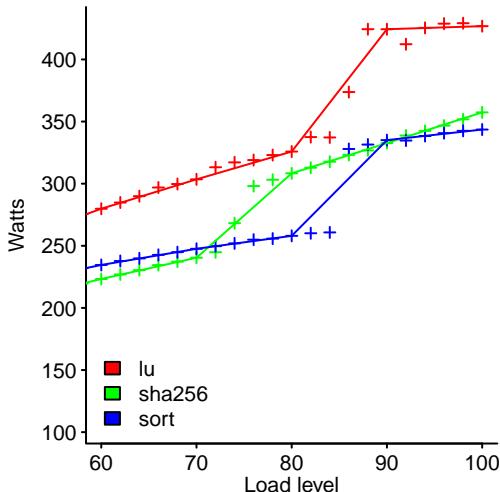


Figure 10.3: A population of items having one of three colors and three strata sampled from it. [code](#)



July 31, 2018  
Figure 10.4: Power consumed by three SERT benchmark programs at various levels of system load; crosses at 2% load intervals, lines based on 10% load intervals. Data kindly provided by Kistowski<sup>1383</sup> [code](#)

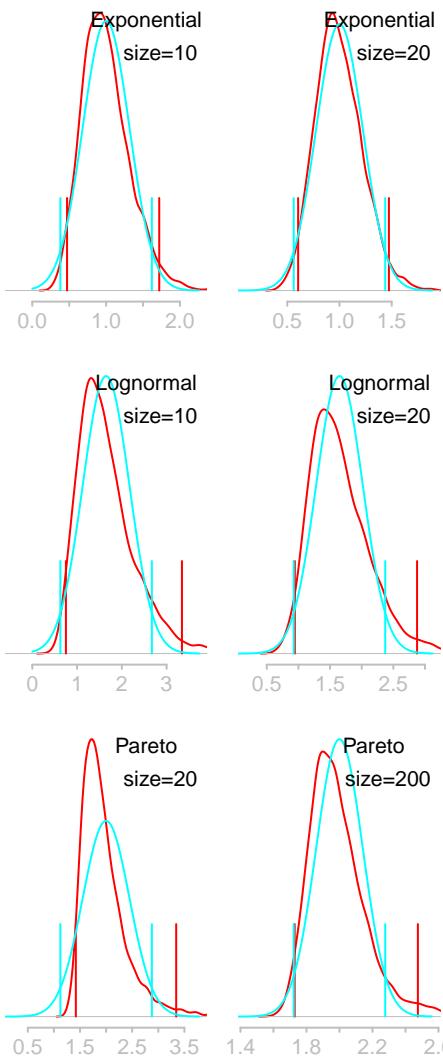


Figure 10.5: Distribution of 4,000 sample means for two sample sizes drawn from exponential (left), lognormal (center) and Pareto (right) distributions, vertical lines are 95% confidence bounds. The blue curve is the Normal distribution predicted by theory. [code](#)

While gathering a representative sample of the population as a whole is a common requirement, sometimes samples having other characteristics are of interest, e.g., being representative of diversity.<sup>987</sup>

### 10.2.1 Sampling error

If the reader is happy to agree that sampling error is an important issue, this section can be skipped. Otherwise, read on and be frightened into agreeing.

The Central Limit Theorem is a statement about the mean value of samples drawn from a population. If the population has a finite variance (power laws with an exponent between zero and two have an infinite variance), then the distribution of sample means converges to a Normal distribution as the sample size,  $N$ , increases (it does not matter which distribution the population has, the distribution of the sample means converges to the Normal).

How quickly does the distribution of sample means converge? The Berry-Esseen theorem provides the best known estimate of the convergence of the distribution of the mean of independent, identically distributed variables to a Normal distribution:

$$|F_n(x) - \Phi(x)| \leq \frac{0.34(\rho + 0.43\sigma^3)}{\sigma^3\sqrt{N}}$$

where:  $F_n$  is the cumulative distribution function of the means,  $\Phi$  the cumulative distribution function of a Normal distribution,  $\rho$  the third moment of  $x$  (and less than infinity),  $N$  the sample size and  $\sigma$  the standard deviation.

The only control parameter available for influencing the error is the number of measurements; the error is proportional to:  $\frac{1}{\sqrt{N}}$ , e.g., to halve the error in the sample mean, the sample size needs to increase by a factor of four.

Experiments have found that for some distributions a sample size of 20 is enough to provide a reasonable approximation to the population mean. However, for other distributions (e.g., those whose tails decay more slowly than exponential) a sample size of over 200 may not be good enough.

Figure 10.5 shows the distribution of mean values for samples drawn from three different distributions (using two sample sizes). The vertical lines are 95% confidence bounds and show that a small increase in sample size is enough for the exponential distribution the bounds to noticeably converge improve (i.e., the confidence bounds of a normal distribution), while a much larger increase in sample size has had no overall impact on the confidence bounds for the Pareto distribution.<sup>iv</sup>

A study by Chen, Chen, Guo, Temam, Wu and Hu<sup>243</sup> measured the performance of programs in the SPEC CPU2006 benchmark using 1,000 sets of input data for each program. As an exercise in sampling let's assume we only have access to three of the possible 1,000 input datasets, what range of execution times might we expect to see from processing just three datasets?

Figure 10.6 was obtained by randomly sampling three items from the population of 1,000 and repeating the process 100 times. The red cross is the sample mean and the vertical grey lines each sample's standard deviation; the blue line is the mean for the population of 1,000 input sets and the green lines the bounds of its standard deviation.

Figure 10.7 shows the distribution of sample means for sample sizes of 3 and 12 items. As expected, the larger samples show less variation in mean value.

The sources of random variability in a sample include the following:

- measurement error caused by the imperfect tools used to make measurements, which can include faults in the programs used to count constructs,
- demographic variability, e.g., particular kinds of programs, or developers working in one location or for one company are measured,
- environmental variability is the sea in which developers swim now or have swum in the past, e.g., company culture or habits acquired from early teachers.

<sup>iv</sup> There will be random fluctuations from the random drawn of the sample.

Figure 10.8 shows the number of commits to glibc<sup>514</sup> for each day of the week, separated out by year. The plot with the large values on the right shows combined counts over all years. The combined plot suggests that most commits occur near the middle of the week, with the number falling off towards the beginning and end of the week. However, the yearly plots rarely show anything like this pattern; is any interpretation of the pattern of commits in the combined plot a just-so story?

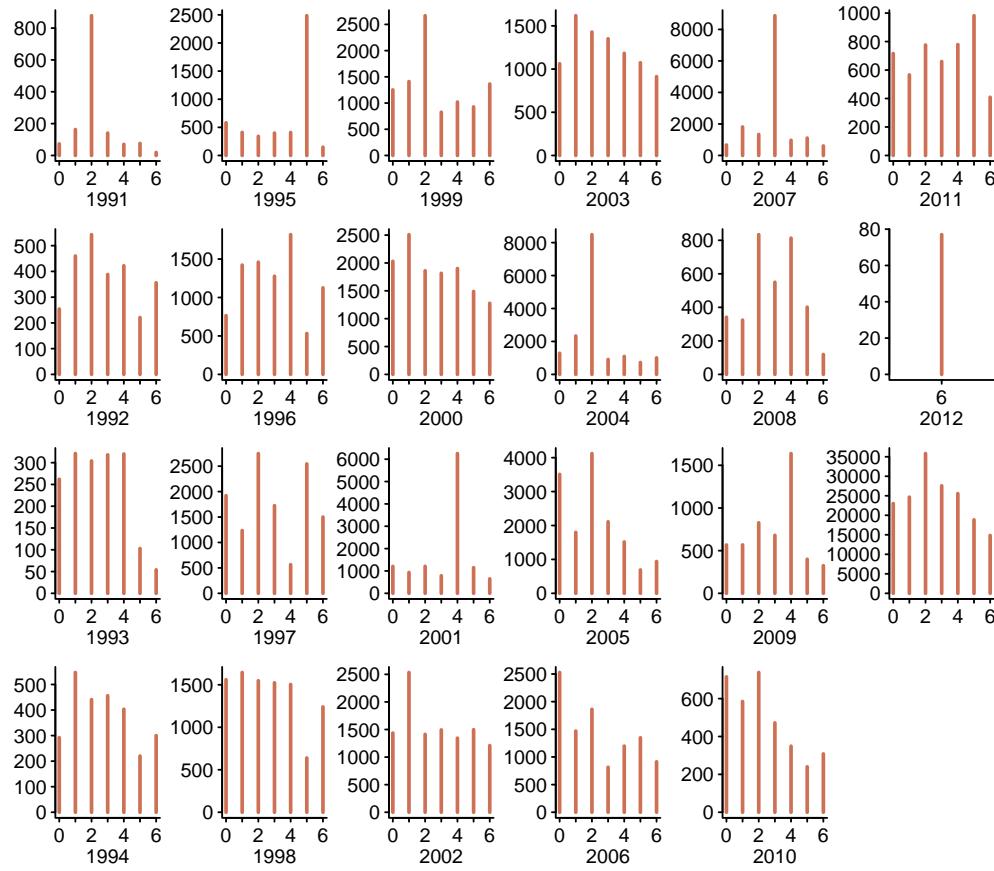


Figure 10.8: Number of commits to glibc for each day of the week, for the years from 1991 to 2012. Data from González-Barahona et al.<sup>514</sup> [code](#)

## 10.3 Describing a sample

A list of values can be overwhelming and compressing many values into a few values, often just one value, is a commonly used approach.<sup>v</sup> The few compressed values are known as *descriptive statistics*. The following are some commonly used algorithms for producing the values that are said to describe a sample:

- a point estimate of a central value and its variability, e.g., mean and standard deviation,
- an equation that closely fits the sample data, such that some condition is met (e.g., minimising mean squared error),
- quartiles cluster measurements based on where values are relative to other values in the sample, e.g., a box-and-whiskers plot such as Figure 8.19.

The mean and standard deviation are the two most commonly used descriptive statistics about a sample. It is incorrect to think that two distributions having the same mean and standard deviation will be very similar; see Figure 10.9.

### 10.3.1 A central location

Perhaps the most widely used single value summary of the values in a sample derives from the idea of a *middle* or *central* location.

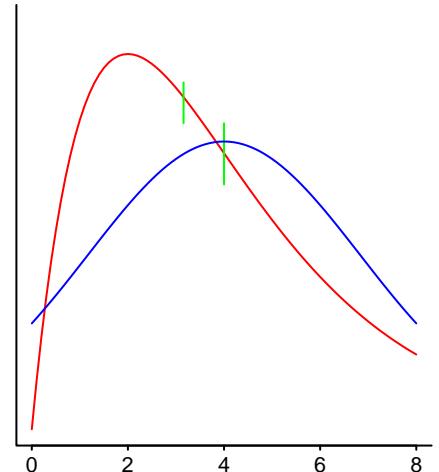


Figure 10.9: A Normal distribution with  $\text{mean}=4$  and  $\text{variance}=8$  and a Chi-squared distribution with four degrees of freedom having the same mean and variance (the vertical lines are at the distributions' median value). [code](#)  
July 31, 2018

<sup>v</sup> Plotting is a technique that often uses all the values and is the primary focus of Chapter 8.

- the mean, obtained by adding together the values in a sample and dividing by the number of values, is perhaps the most commonly used central location,
- the *median* is obtained by sorting the  $N$  values into numerical order and selecting the value of the  $\frac{N+1}{2}$ th element (if  $N$  is even the average of the middle two values is used),
- the *mode* is the value most likely to be sampled (the mode function is unrelated to the statistical algorithm of that name, it returns the type or storage mode of an object). The modeest package contains functions for estimating various kinds of mode.

For symmetric distributions the values of the mean, median and mode are equal, and for asymmetric distributions the three values can be very different.

For any sample values drawn from a unimodal distribution, the difference between the median and mean is less than or equal to  $\sqrt{0.6}\sigma$ , and for other distributions less than  $\sigma$ .

The difference between the median and mode is less than or equal to  $\sqrt{3}\sigma$ .

Unless the sample distribution is symmetric, it is not possible to sum multiple modes, e.g., cost estimates. For nonsymmetric distributions, adding underestimates the true value, e.g., for a gamma distribution the mean is  $k\theta$  and the mode is  $(k - 1)\theta$ , where  $k$  and  $\theta$  describe the distribution.

Figure 10.10: Density plot of execution time of 1,000 input data sets, with lines marking the mean, median and mode. Data kindly supplied by Chen.<sup>243</sup> [code](#)

Figure 10.10 shows the distribution of execution times of the 1,000 input data sets from Chen et al.<sup>243</sup> If we are interested in an estimate of the execution time of a randomly chosen input data set, it makes sense to use the median value, the point that equally divides the number of input data sets. If we are interested in an estimate of the execution time most likely to be encountered, the value to use is the mode.

Some distributions have such fat tails that the mean is infinite, e.g., the Cauchy distribution. In practice the regularity with which very large values occur causes the mean value of a sample to jump around erratically, as new measurements are added to it. A distribution that does not have a finite mean may still have a median. This is because the median is not affected by extreme values in the way the mean is, so any extreme values that do appear in a sample do not prevent the median converging to a fixed value.

The well-known algorithms for calculating the mean and standard deviation of a sample require that each value be independent of the others. When a sequence of values is serially correlated, i.e., the value of a measurement is related to the value of the one or more immediately previous measurements, the calculated mean and standard deviation is biased. In the case of the mean, the uncertainty in its value grows for positive correlation and decreases for negative correlation. The upper plot in Figure 10.11 shows the fraction of this change for various sample sizes; it is based on an AR(1) model, where each value correlates with the immediately preceding value by an amount given in the legends on the right of the plot (see Section 11.10 for details of AR models). A positive correlation causes the ratio of the sample standard deviation, relative to the population standard deviation, to be underestimated, while a negative correlation causes it to be overestimated (the lower plot in Figure 10.11 shows the fraction of this change).

The sandwich package supports the calculation of various error measures caused by serial correlation (e.g., the lrvar function calculates the error in the long term mean of a series).

**Compositional data:** The individual components of compositional data are correlated and the mean of each component cannot be calculated independently of the other components. The mean function in the compositions package calculates the mean of compositional data.

Several methods of calculating the variance and standard deviation of compositional data have been proposed. The compositions package supports the mvar function, which calculates what is known as the *total variance* (or *generalized variance*) and the msd function which calculates the *metric standard deviation* (both return single values). The variation matrix includes information about the relationship between pairs of components and is returned by the variation function, for instance (see Figure 5.29 for details): [code](#)

	Design_Sched	Code_Sched	Systest_Sched	Acctest_Sched
Design_Sched	0.0000000	0.2278511	0.3534392	0.3340407
Code_Sched	0.2278511	0.0000000	0.4233143	0.4523761
Systest_Sched	0.3534392	0.4233143	0.0000000	0.4787860
Acctest_Sched	0.3340407	0.4523761	0.4787860	0.0000000

Figure 10.11: Impact of serial correlation, AR(1) in this example, on the calculated mean (upper) and standard deviation (lower) of a sample (the legends specify the amount of serial correlation). [code](#)

**Circular data:** Figure 11.76 illustrates a calculation of the mean of values drawn from a circular distribution.

### 10.3.2 Sensitivity of central location algorithms

Samples sometimes include extreme values (which may or may not be the result of noise). The percentage of observations that can cause a statistical estimator to produce an arbitrarily large (positive or negative) value is known as the *breakdown point*.

The breakdown point for the mean is proportional to  $\frac{1}{N}$ , i.e., no matter how many observations are made it only takes one extreme value to produce a completely spurious result for the mean; the mean has the smallest breakdown point it is possible to have.

At the other end of the scale the median has a breakdown point of 0.5 (i.e., half of the measurements can have extreme value without affecting the value of the result) and for this reason the median is often recommended, over the mean, when measurements values are known to be very noisy. However, the median cannot be recommended for universal use because there are situations where it does not perform as well as the mean. For instance, when values are drawn from a discrete distribution whose mean is roughly half-way between measurable points and the sample includes duplicate values, then most samples will have a median value slightly larger/smaller than the actual mean, i.e., the median is not evenly distributed across possible values in the way the Central limit theorem says that the mean is distributed; see Figure 10.12.

The probability of an outlier occurring depends on the reliability of the measurement process and the characteristics of the population being sampled from. The following two techniques are robust in the presence of extreme values and often used when a single mean value is involved:

- *trimmed mean* removes a percentage of the largest and smallest values before calculating the mean of the remaining values (it has been found that 20% is a good value for general use). The `mean` function includes a `trim` argument for specifying the percentage to be trimmed.
- *winsorized mean* replaces rather than remove values. The values of the lowest X% are replaced with the lowest value that is just not within percentage, and the values of the highest X% are replaced with the highest value just not within this percentage; the Winsorized mean is calculated using the updated list of values. The `psych` package contains functions that calculate various quantities using the Winsorizing algorithm.

The Trimmed and Winsorized means do not produce biased results when applied to samples drawn from a population having a symmetric distribution.

### 10.3.3 Geometric mean

The *geometric mean* is defined as:

$$\text{Mean}_g = \left( \prod X \right)^{\frac{1}{N}}$$

e.g., the geometric mean of 10, 100, 1000 is  $(10 \cdot 100 \cdot 1000)^{\frac{1}{3}} \rightarrow 100$ .

The geometric mean is preferred to the arithmetic mean when ranking ratios or normalised data (which is a kind of ratio) because it gives consistent results.

Consider the (invented) benchmark performance of the three systems in Table 10.1. Treating  $a$  as the base performance, what is the relative performance improvement of  $b$  and  $c$ ?

If the arithmetic mean is used, the performance ranking of  $b$  and  $c$ , relative to  $a$ , depends on whether the calculation used is a ratio of the means or the mean of the ratios. The arithmetic means of the benchmark performance for each system is 50.5, 53.5 and 53: the ratio of these mean values is listed in the fourth column; the individual benchmark ratios for  $a$  and  $b$  are:  $\frac{2}{1}$  and  $\frac{105}{100}$ , and for  $a$  and  $c$ :  $\frac{3}{1}$  and  $\frac{103}{100}$ . The mean of these ratios is listed in the fifth column. Comparing columns four and five shows that the ranking of  $b$  and  $c$  depends on the method of calculation.

If the geometric mean is used, the relative order of the final ratio is not order dependent.

Sometimes the arithmetic and geometric means produce the same benchmark rankings, e.g., a benchmark<sup>398</sup> of eight Intel IA32 processors used the arithmetic mean of ratios to compare results, the results from using the geometric means was not large enough to affect the relative

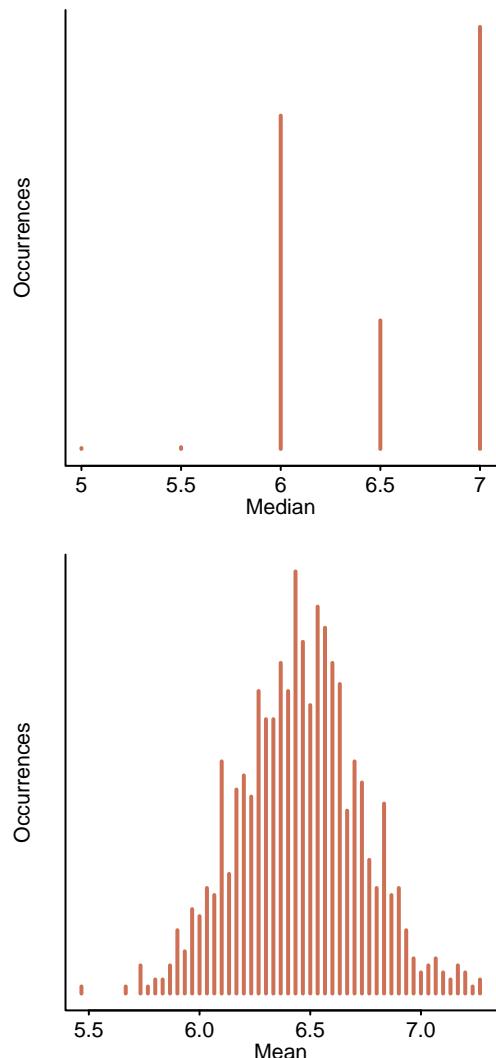


Figure 10.12: Occurrence of sample median and mean values for 1,000 samples drawn from a binomial distribution.  
code

system	integer	float	ratio of means	mean of ratios	geometric mean
a	1	100			10
b	2	105	$\frac{53.5}{50.5} \rightarrow 1.0594$	mean(2/1+105/100) -> 3.05	14.49
c	3	103	$\frac{53}{50.5} \rightarrow 1.0495$	mean(3/1+103/100) -> 4.03	17.58

Table 10.1: Invented benchmark performance measurements of three systems and various methods of calculating relative performance. The relative performance of *b* and *c* depends on which mean is used.

ranking of processors for a given performance characteristic (see [benchmark/powervperfasp-los2011.R](#)).

The Geometric mean might be used when values cover several orders of magnitude, e.g., a geometric or logarithmic series (e.g., 2, 4, 8, 16, 32, 64, ...)

Techniques for calculating the geometric mean include the expression `exp(mean(log(x)))` and the `geometric.mean` function in the `psych` package.

### 10.3.4 Harmonic mean

The *harmonic mean* is used to find the "average" of a set of ratios or proportions; it is defined as:

$$Mean_h = \frac{N}{\sum X^{-1}}$$

e.g., the harmonic mean of 1, 2, 3, 4, 5 is  $\frac{5}{\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}} \rightarrow 2.189781$

When there are two values the formula becomes:

$$Mean_h = 2 \frac{x \cdot y}{x + y}$$

which has the same form as the  $F_1$  score, or F-measure, used in information retrieval to combine the precision and recall:

$$F_1 = (1 + \beta^2) \frac{precision \cdot recall}{\beta^2 precision + recall}$$

Two ways of calculating the harmonic mean include the code `1/mean(1/x)` and the `harmonic.mean` function in the `psych` package.

### 10.3.5 Contaminated distributions

That a sample, or the error present in the measurements it contains, can be fitted to a Normal distribution is a common assumption that often goes unquestioned. Textbooks are filled with techniques that only exhibit the cited desirable attributes when the Normality assumption holds. There is also the lure of analytic solutions to problems, which again are often only available when the Normality assumption holds.

Even when sample values appear to be drawn from a Normal distribution, a small percentage of contaminated values can have a dramatic effect on the value returned by a statistical algorithm.

The Contaminated Normal distribution is a mixture of values drawn from two Normal distributions both having the same mean, but with 10% of the data points drawn from a distribution whose standard deviation is five times greater than the other. Figure 10.13 shows the sample density where 10,000 values are drawn from a Normal distribution and 10,000 values from a Contaminated Normal distribution; visually they seem very similar (see [statistics/contam-norm.R](#) to learn the color used to plot each sample).

This contaminated Normal distribution has a standard deviation that is more than three times greater than the Normal distribution from which 90% of them are drawn. This illustrates that a Normal distribution contamination by just 10% of values from another distribution can appear to be Normal, but have very different descriptive statistics.

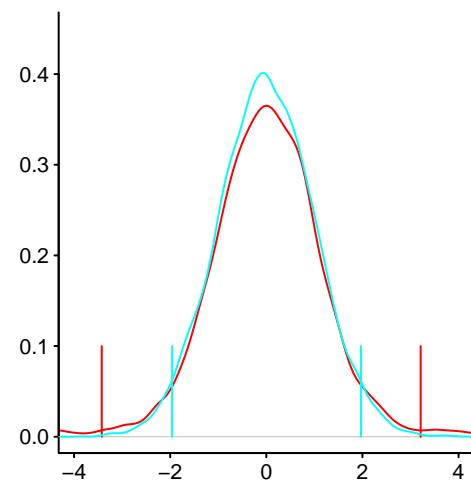


Figure 10.13: A contaminated normal, values drawn from two normal distributions with 10% of values drawn from a distribution having a standard deviation five times greater than the other. [code](#)

There are a number of tests for estimating whether sample values are drawn from a Normal distribution. The Shapiro-Wilk test (the `shapiro.test` function), the Kolmogorov-Smirnov Test (the `ks.test` function)<sup>vi</sup> and the Anderson-Darling test are common suggested techniques. A comparison of four normality tests<sup>1147</sup> found the Shapiro-Wilk test to be the most powerful normality test; see Figure 9.9.

Recommendation: Use the Shapiro-Wilk test for testing whether sample values are drawn from a Normal distribution.

When a data set contains very few values, even the Shapiro-Wilk test may fail to detect (e.g., p-value < 0.05) that sample values are not drawn from a Normal distribution. In the case of the Contaminated Normal distribution, samples containing only 10 values are considered to have a Normal distribution in around 30% of cases (based on p-value > 0.05), with the percentage dropping to 10% for samples containing 20 values.

For an in depth analysis of the potential problems that outliers, skewed distributions and fat tails can cause see "Introduction to Robust Estimation & Hypothesis Testing" by Rand Wilcox.<sup>1451</sup>

## 10.4 Statistical error

The outputs from a statistical technique generally includes probabilities and it is the responsibility of the user of the technique to decide the cut-off probability below/above which an event is considered to have/not occurred.

The two kinds of error that can be made are:

- treating a hypothesis as true when it is actually false (known as a *False positive*), the technical statistical term is making a *Type I* error;  $P(\text{TypeIerror} = P(\text{Reject } H_0 | H_0 \text{ true}))$ ,
- treating a hypothesis as false when it is actually true (known as a *False negative*), the technical statistical term is making a *Type II* error;  $P(\text{TypeIIerror} = P(\text{Do not reject } H_0 | H_A \text{ true}))$ , where  $H_A$  is an alternative hypothesis).

		Decision made	
		Reject $H$	Fail to reject $H$
Actual	$H$ true	Type I error	Correct
	$H$ false	Correct	Type II error

Table 10.2: The four combinations of circumstances and their outcomes in hypothesis testing.

The consequences of making an error will depend on the perspective of those affected by the outcome of the decision. For instance, consider the consequences of a manager's decision on whether to invest more time and money testing the reliability of a software system; an incorrect decision can lead to more than losing the original investment (i.e., losing market share to a competitor), but the likely bearer of the loss will depend on the vendor/customer relationship world in which the decision plays out, as Table 10.3 illustrates:

		Decision made	
		Finish testing	Do more testing
Actual	More testing needed	Customer loss	Ok
	Testing is sufficient	Ok	Vendor loss

Table 10.3: Finish/do more testing decision and likely outcome based on who incurs the loss.

### 10.4.1 Hypothesis testing

A hypothesis is an unverified explanation of why things are the way they are. Hypothesis testing is the process of collecting and evaluating evidence that may or may not be consistent with the hypothesis, i.e., positive and negative testing.<sup>vii</sup> Once enough evidence consistent

<sup>vi</sup> Both included in the R base system.

<sup>vii</sup> Gigerenzer<sup>494</sup> is a readable discussion of how people cope with uncertainty.

with the hypothesis has been collected people may feel confident enough to start referring to it as a theory or law.

The most commonly used hypothesis testing technique is based on what is known as the null hypothesis<sup>viii</sup>, which works as follows:

- a hypothesis,  $H$ , having testable prediction(s) is stated,
- an experiment to test the prediction(s) is performed, producing data  $D$ ,
- assuming the hypothesis is true, calculate the probability of obtaining the data produced by the experiment. The calculation made is:  $P(D|H)$ ; that is the probability of obtaining the data  $D$  assuming that the hypothesis  $H$  is true.

If the calculated probability is less than or equal to some prespecified value, the hypothesis is rejected, otherwise it is said that *the null hypothesis has not been rejected* (since the result of the experiment is not conclusive evidence that the null hypothesis is true).

Expressed in code, the null hypothesis testing algorithm is as follows:

```
void null_hypothesis_test(void *result_data, float p_value)
{
    // H is set by reality, only accessed by running experiments
    if (probability_of_seeing_data_when_H_true(result_data) < p_value ||
        !H)
        printf("Willing to assume that H is false\n");
    else
        printf("H might be true\n");
}

null_hypothesis_test(run_experiment(), 0.05);
```

A test statistic is said to be *statistically significant* when it allows the null hypothesis to be rejected. The phrase "statistically significant" is often shortened to just "significant", a word whose common usage meaning is very different from its statistical one; this shortened usage is likely to be misconstrued when the audience is not aware that the statistical definition is being used and assumes the word is being used in its everyday meaning sense.

Statistical significance can roughly be treated as meaning *likelihood of occurring by chance*, it does not mean the results highlighted by the statistical analysis have any practical significance, i.e., the magnitude of the pattern detected may still be so small as to make it useless for practical applications.

Running one experiment that produces a surprisingly high/low p-value is a step in the process of increasing everybody's confidence that a hypothesis is true/false.

Replication of the results (i.e., repeating the experiment and getting the very similar measurements) provides evidence that the first experiment was not a chance effect; a further boost in confidence. Replication by others, who independently set up and run an experiment, is the ideal replication (it reduces the possibility that unknown effects specific to a person or group influenced the outcome); an even larger boost in confidence.

There is a great deal of confusion surrounding how the results from a null hypothesis test should be interpreted. Studies have found<sup>496</sup> that people (incorrectly) think that one or more of the following statements apply:

- *Replication fallacy*: The level of significance measures the confidence that the results of an experiment would be repeatable under the conditions described. This is equivalent to saying:  $P(D|H) == 1 - P(D)$ , and would apply if the hypothesis was indeed true.
- the significance level represents the probability of the null hypothesis being true. This is equivalent to saying:  $P(D|H) == P(H|D)$ .

The Bayesian approach to hypothesis testing is growing in popularity and works as follows:

---

<sup>viii</sup> As the market leader in hypothesis testing techniques, over many decades, this technique attracts regular criticism.<sup>271</sup> The criticism is invariably founded on widespread misuse of the null hypothesis ritual; misuse is the fate of all widely used techniques.

- two hypotheses,  $H_1$  and  $H_2$ , having testable prediction(s) are stated (the second hypothesis may just be that  $H_1$  is false),
- a non-zero probability is stated for the hypotheses being true,  $P(H_1)$  and  $P(H_2)$ , known as the *prior* probabilities,
- an experiment to test the prediction(s) is performed (producing data  $D$ ),
- update the previously estimated probability that  $H_1$  and  $H_2$  are true. The calculation uses Bayes theorem, which for  $H_1$  is:

$$P(H_1|D) = \frac{P(H_1)P(D|H_1)}{P(H_1)P(D|H_1) + P(H_2)P(D|H_2)}$$

The updated prior probability, on the basis of the experimental data, is known as the *posterior probability* of the hypothesis being true.

### 10.4.2 p-value

In a randomized experiment the *p-value* is the probability that random variation alone produces a test statistic as extreme or more extreme than the one observed.

In a commercial environment the choice of p-value should be regarded as an input parameter to a risk assessment comparing the costs and benefits of all envisioned possibilities.

In many social sciences the probability of the Null hypothesis being true must be less than 0.05 (i.e., 5%, or slightly greater than  $2\sigma$ ),<sup>ix</sup> while in civil engineering a paper describing a new building technique that created structures having a 1-in-20 chance of collapsing would not be considered acceptable. High energy physics requires a p-value below  $5\sigma \rightarrow 5.7 \cdot 10^{-7}$ , before the discovery of a new particle is accepted.

As sample size increases, p-values will always become smaller. For instance, some aspect of flipping a coin may very slightly favour heads and given enough coin flips a sufficiently small p-value, for the hypothesis that the coin is not a fair one, will be obtained. There is no procedure for adjusting p-values for hypothesis analyses using very large amounts of data.

When lots of measurement data about many variables is available it is possible to go on a fishing expedition, looking for relationships between variables.<sup>1165</sup> The probability of finding one significant result when comparing  $n$  pairs of variables, using a p-value of 0.05, is  $1 - (1 - 0.05)^n$  (which is 0.4 when  $n = 10$ ). When multiple comparisons are made the base p-value needs to be adjusted to take account of the increased probability of treating noise as signal.

Perhaps the most common technique is the *Bonferroni correction*, which divides the base p-value by the number of tests performed. In the above example, the base p-value would be adjusted from 0.05 to  $\frac{0.05}{10} \rightarrow 0.005$ , for each of the ten tests.

The `p.adjust` function supports p-value adjustment using a variety of different techniques.

Researchers know their work only has a chance of being accepted for publication if the reported results have p-values below the journal's cut-off value. Given the use of published paper counts as a measure of academic performance, there is an incentive for researchers to run many slightly different experiments to find a combination that produces a sufficiently low p-value that the work can be written up and submitted for publication<sup>714</sup> (a process known as *p-hacking*).<sup>x</sup> One consequence of only publishing papers containing studies achieving a minimum p-value is that many of the results are likely to be false (while a theoretical analysis suggests most are false,<sup>666</sup> an empirical analysis suggests around 14% of false positives for medical research<sup>680</sup>).

### 10.4.3 Confidence intervals

Many statistical techniques return a single number, a point value. What makes this number so special, would a value close to this number be almost as good an answer? If an extra measurement was made and added to the sample, is this number likely to dramatically change;

<sup>ix</sup> Journals with high impact ratings can be more choosy and some specify a p-value of 0.01.

<sup>x</sup> Which commercial company would not willing add warts to software to keep an important customer happy?

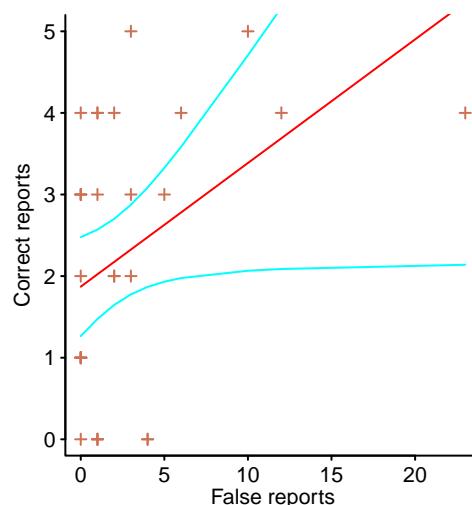


Figure 10.14: Regression model (red line; pvalue=0.02) fitted to the number of correct/false security code review reports made by 30 professionals; blue lines are 95% confidence intervals. Data from Edmundson et al.<sup>375</sup> [code](#)

what is one measurement were excluded from the sample, how much would that change the answer?

A confidence interval is an upper and lower bound on the result of a statistical technique. A common choice is the 95% confidence bound, default value used by many R library functions.

The numeric values associated with a *confidence interval* can be translated into visual form by including them in a plot. Figure 10.14 illustrates how confidence intervals provide an easier to digest insight into the uncertainty of a fitted regression model, compared to the single number that is the p-value. The red line shows a fitted regression model, whose predictor has a p-value of 0.02, with 95% confidence intervals in blue, showing how wide a range of lines could be said to fit the sample just as well.

A confidence interval is a random variable, it depends on the sample drawn. If many 95% confidence intervals are obtained (one from each of many samples), the true fitted model is expected to be included in this set of intervals 95% of the time (it is a common mistake to think that the confidence interval of one sample has this property). The probability that the next sample will be within the 95% confidence interval of the current sample, for a Normal distribution, is 84% or around 5 out of 6.<sup>301</sup>

A closed form formula for calculating confidence intervals is only known for a few cases, e.g., the mean of samples drawn from a Normal distribution, for a Binomial distribution a variety of different approximations have been proposed.<sup>1093</sup>

Built-in support for calculating confidence intervals, in R packages, is sporadic. A Monte Carlo algorithm can be used to calculate a confidence interval from the sample, e.g., the bootstrap. This approach has the advantage that it is not necessary to assume that sample values are drawn from any particular distribution. Figure 10.15 was created by fitting many models, via bootstrapping, and using color to indicate density of fitted regression lines.

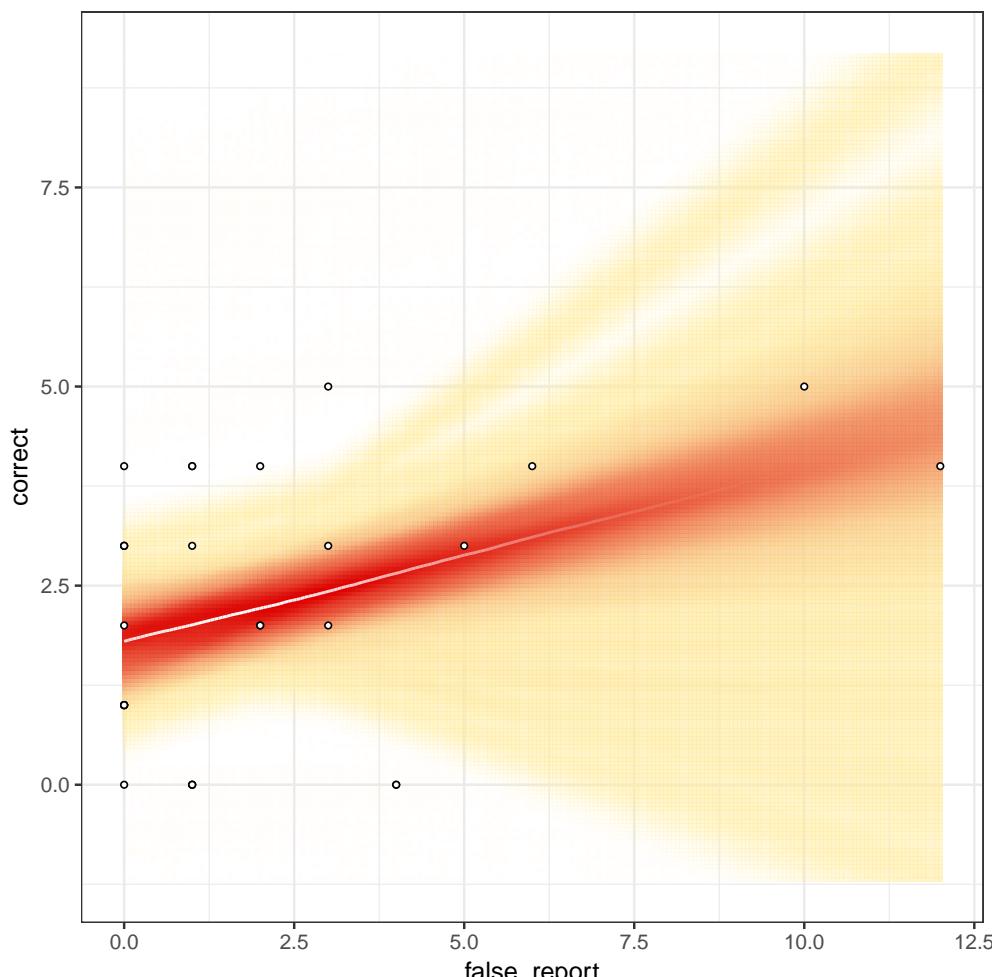


Figure 10.15: Bootstrapped regression lines fitted to random samples of the number of correct/false security code review reports made by 30 professionals. Data from Edmundson et al.<sup>375</sup> [code](#)

### 10.4.4 The bootstrap

The bootstrap is a general technique for answering questions about the uncertainty in the estimate of a statistic calculated from a sample (e.g., calculating a confidence interval or standard error).<sup>612</sup> Bootstrap techniques operate on a sample drawn from a population and cannot extract information about the population that is not contained in the sample, e.g., if the population contains reds and greens and a sample only contains reds, then the bootstrap will not provide any information about the greens.

Bootstrapping computing is applied to the process by which systems start themselves. In statistics, it is used to describe a process where new samples are created from an existing sample; the term *resampling* is sometimes used.

Estimating the confidence interval for the mean value of a sample is a good example of the basic bootstrap algorithm; the steps involved are as follows:

- create a sample by randomly drawing items from the original sample. Usually the items are selected with replacement (i.e., an item can be selected multiple times). When items are selected without replacement (i.e., can only be selected once), the term *jackknife* is used,
- obtain the mean value of the created sample,
- iterate, say, 5,000 times.
- analyse the 5,000 mean values to obtain the lowest and highest 2.5%. The 95% confidence interval for the mean of the original sample is calculated from this lowest/highest band (several algorithms, giving slightly different answers, are available);

The `boot` package support common bootstrap operations, including `boot.ci` for obtaining a confidence interval from a bootstrap sample.

The distribution of the sample from which the bootstrap algorithm draws values is known as the *empirical distribution*.

The *bootstrap distribution* contains  $m^n$  possible samples, when sampling with replacement from  $m$  possible items to create samples containing  $n$  items; when the order of items does not matter, there are  $\binom{2m-1}{n}$  possible samples (a much smaller number).

The same bootstrap procedure can be applied to obtain confidence intervals on a wide range of metrics. Figure 10.16 shows confidence intervals for kernel density in Figure 8.14 and was produced by `sm.density`, in the `sm` package, using the following code:

```
library("sm")

res_sample=sample(cint$result, size=1000) # generate 1000 samples

sm.density(res_sample, h=4, col=point_col, display="se", rugplot=FALSE,
           ylim=c(0, 0.03),
           xlab="SPECint Result", ylab="Density\n")
```

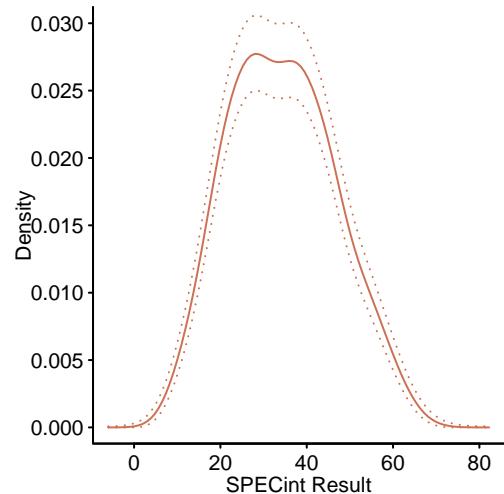


Figure 10.16: Kernel density plot, with 95% confidence interval, of the number of computers having the same SPECint result. Data from SPEC.<sup>1282</sup> code

### 10.4.5 Permutation tests

For small sample sizes, computers are fast enough for it to be practical to calculate a statistic (e.g., the mean) for all possible permutations of the items in a sample. This kind of test is known as a *permutation test*.

Some techniques designed for manual implementation (e.g., Student's t-test) are approximations to the exact answer produced by a permutation test.

Permutation tests do not have any preconditions on the distribution of the sample, other than it is representative of the population, and produce an exact answer.

The `coin` package contains infrastructure for creating permutation tests and functions that perform common tasks (the names of these functions are derived from the names of the tests designed for manual implementation, e.g., `spearman_test` and `wilcox_test`).

The following is a permutation test of whether the professional experience of the two samples of subjects shown in Figure 8.1 are likely to have different mean values:

```
library("coin")

# The default is alternative="two.sided",
# an option not currently listed in the Arguments section.
oneway_test(experience ~ as.factor(language), data=Perl_PHP,
            distribution="exact")
```

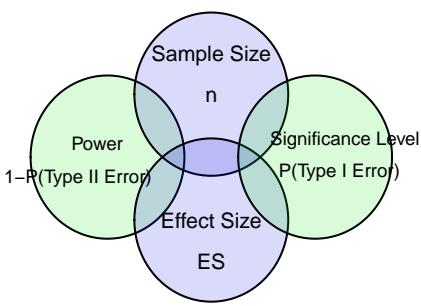


Figure 10.17: The four related quantities in the design of experiments. [code](#)

## 10.5 Effect-size

*Effect-size* is the degree to which the characteristic of interest is present in the population (from which a sample is drawn), e.g., if we are interested in the difference in the mean performance of developers before and after attending a training course, how big is the difference (answering this question may be the reason for obtaining measurements)?

The question to ask about a given effect-size is: “Does it matter?” The larger the effect-size the more likely it is to be of interest in practice; in some cases a small effect-size may be of interest (e.g., a small difference multiplied over a large population can have a large impact), while in other cases only a large effect-size is of interest (e.g., when the population is small a large effect-size could be needed to have a large impact).<sup>xi</sup>

Smaller effect-sizes are likely to be more costly to detect because more measurements are needed to isolate small effect-sizes compared to larger ones.

Figure 10.18 shows how percentage differences in the presence of a condition in a population can have a dramatic effect on the false positive rate (in red), for the same statistical power and p-value.

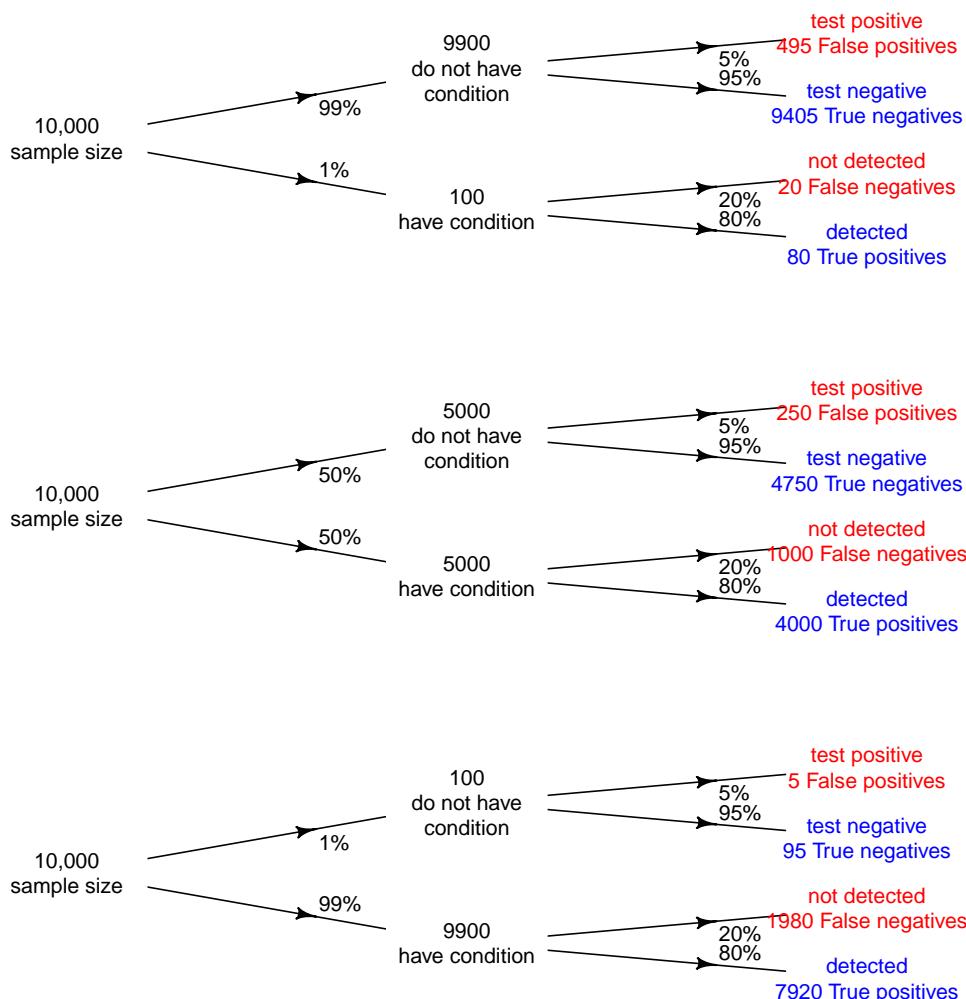


Figure 10.18: Examples of the impact of population prevalence, statistical power and p-value on number of false positives and false negatives. [code](#)

<sup>xi</sup> Statistical books<sup>270</sup> and papers sometimes concern themselves with questions of where to draw the lines that delimit large/medium/small effect-sizes; an approach that might be applicable when researchers are more interested in publishing papers than making useful discoveries.

Methods for calculating effect-size depend on how data is being analysed<sup>386</sup> and include the following:

- correlation, e.g., the Pearson correlation coefficient, is a measure of effect-size,
- combining information on the mean and standard deviation of two samples into a single value. For instance, Cohen's  $d$  is one measure used when the samples have similar standard deviations and is given by (other approaches adjust the calculation of the standard deviation):  $d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$ .

Figure 10.19 illustrates how differences in mean and standard deviation, of two distributions, result in a given Cohen's  $d$ ,

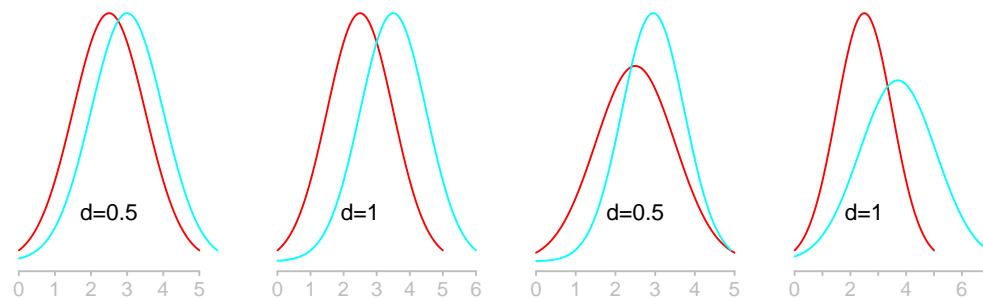


Figure 10.19: Visualization of Cohen's  $d$  for two normal distributions having different means and the same standard deviation (two left) and both different (right). [code](#)

- odds ratio, that is the ratio of the odds (i.e.,  $\frac{p}{1-p}$ ) of an event occurring in one sample divided by the ratio of the same event occurring in the other sample (perhaps a control group),
- other methods are discussed where applicable.

## 10.6 Statistical power

If an effect exists and an experiment is performed to measure it, what is the likelihood that the effect will be detected? The numeric answer to this question is known as *statistical power*, of the experiment. The *power* of a statistical test is its ability to detect a difference when one actually exists in the data. Failing to detect an effect when one exists is known as making a *Type II error*, or more commonly as a false negative ( $\beta$  is commonly used to denote the Type II error rate). Techniques for reducing Type II errors include:

- being willing to accept a larger *Type I error* ( $\alpha$  is commonly used to denote the Type I error rate),
- sample from a population that has a higher probability of containing the sought after characteristics. For instance, Vasa<sup>1394</sup> excluded releases with less than 30 changed classes in a study of class change dynamics. If a subset of a population is selected to maximise detection rate, care must be taken to ensure that any statement of statistical power refers to the subset population, not the larger population from which it was subsetted,
- increasing the number of measurements made.

The area of the unknown distribution excluding  $\beta$  (i.e.,  $1 - \beta$ ) is known as the power of the test.

A power of 80% is often quoted<sup>270</sup> as being an acceptable lower limit of a test having a high power, just like 5% is often quoted as an acceptable significance level in many disciplines.

As an example, Figure 10.20 shows the distribution of measurements in two populations: X (red) and Y (green) (e.g., the time taken to execute all possible programs, with all possible input, on two different computers). The upper and middle plot only differ in mean value, while the middle and lower plot only differ in standard deviation. The false positive rate,  $\alpha$ , is shaded in red, and the false negative rate,  $\beta$ , in green. The two rates are connected in that increasing one decreases the other, and vice versa.

Measuring an entire population is not usually practical, (e.g., all programs and over all input data) will not be available; a sample of the population is measured.

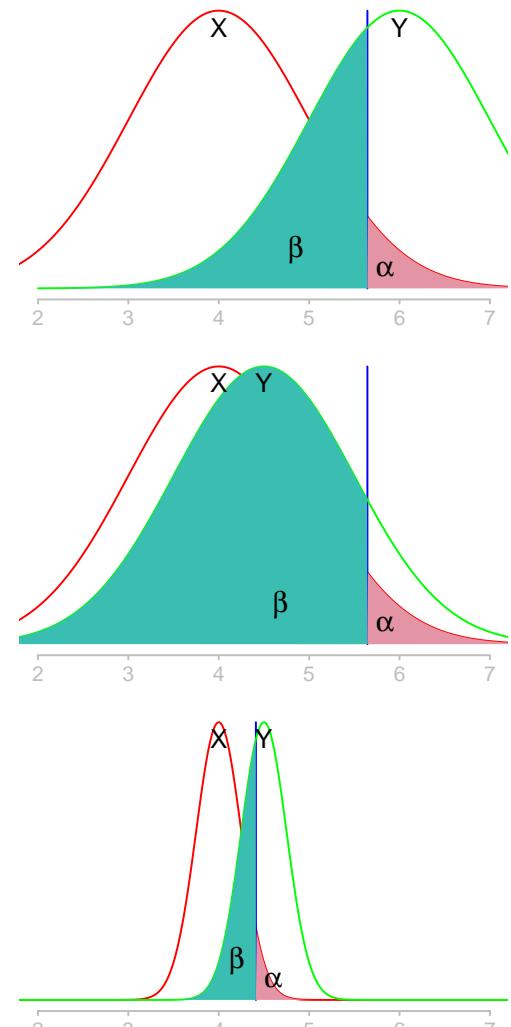


Figure 10.20: The impact of differences in mean and standard deviation on the overlap between two populations ( $\alpha$ : probability of making a false positive error, and  $\beta$ : probability of making a false negative error). [code](#)

When there is a large overlap between populations (middle plot), most of the measurements in the Y sample may have values that suggest that were drawn from the X population. There is less overlap in the upper and lower plots and sample values are more likely to appear to be drawn from different distributions.

In the upper plot, the larger difference in the population mean makes it more likely that sample measurements from Y will have values that are more likely to appear to be drawn from a different distribution than samples from X. In the lower plot, there is less overlap because of the smaller population standard deviations.

If there is a need to find out whether an effect exists (e.g., one computer is faster than another or a new algorithm uses less memory), the first question is whether the effect is likely to be detected using the available resources (e.g., time and effort needed to obtain a measurement sample). A statistical power calculation enables the tradeoffs between sample size and probability of detecting an effect (assuming a given population mean, standard deviation and amount of difference between two or more samples) to be analysed.

The Monte Carlo simulation can be used to obtain an estimate of the likelihood that a particular test will detect an effect in a sample of the population. The algorithm works by simulating the experiment under consideration by, obtaining samples from the population(s) that are thought to exist and performing the analysis on each sample, counting each success/failure to detect the difference.

The following code creates two populations and then compares two samples drawn from these populations. The user written function `some_test_statistic` compares two samples and returns the probability that they have some property (`statistics/boot-power.R` contains an example that checks for a difference in mean value between samples drawn from two populations, see Figure 10.21):

```
boot_power=function(pop_1, pop_2, sample_size, test_stat, alpha=0.05)
{
  num_samples=5000 # Number of times to run the 'experiment'.
  results=sapply(1:num_samples, function(X)
  {
    sample_1=sample(pop_1, size=sample_size, replace=TRUE)
    sample_2=sample(pop_2, size=sample_size, replace=TRUE)
    return(test_stat(sample_1, sample_2, alpha))
  })

  return(sum(results<alpha)/num_samples) # percentage detected
}

# Create two slightly different populations.
population_1=rnorm(100000, mean=0, sd=1)
population_2=rnorm(100000, mean=0+0.5, sd=1)

boot_power=function(population_1, population_2, 20,
                    some_test_statistic, alpha=0.05)
```

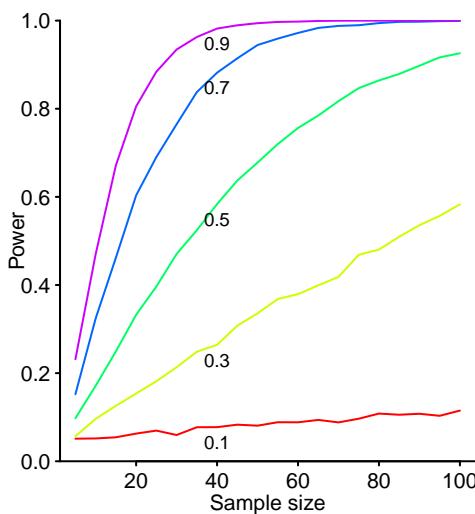


Figure 10.21: The power of a statistical test at detecting that a difference exists between the mean values of samples drawn from two populations, both having a Normal distribution; actual mean difference adjacent to colored line. [code](#)

In the case where both populations have a Normal distribution, an analytic solution is available for calculating the statistical power of using the t-test to detect a difference in two sample means (available in the `power.t.test` function). However, except for a few special cases, analytic solutions are not known. The lack of an analytic solution is not a problem if a computer is available, a Monte Carlo approach can be used.

Figure 10.21 shows a plot of the results of a Monte Carlo simulation testing for a difference in the mean of two samples drawn from two populations (randomly generated, using `rnorm`, to have various differences in their mean), using various sample sizes (see `statistics/response-power.R` for the values returned by the analytic solution).

Obtaining good enough accuracy from a power analysis requires a good approximation of the characteristics of the population distribution. This information might come from the results from the analysis of related measurements, a preliminary study or theory.

A study by Syed, Robinson and Williams<sup>1332</sup> investigated variations in the number of failures experienced when using the Firefox browser, at different processor speeds, system memory and hard disc sizes. A total of 11 known mistakes causing intermittent failure (four of these were not experienced) and nine different hardware configurations were selected. The conditions expected to experience each fault were created and Firefox was executed 10 times for

each hardware configuration. Table 10.4 shows the number of each fault experienced in each hardware configuration.

Mhz-Mb-Gb	124750	380417	410075	396863	494116	264562	332330
667-128-2.5	4	10	6	5	2	3	5
667-256-10	4	8	8	6	4	3	8
667-1000-2.5	4	7	3	4	3	1	8
1000-128-10	3	10	3	6	0	1	1
1000-256-2.5	3	9	0	6	0	1	2
1000-1000-10	2	9	4	5	0	0	1
2000-128-2.5	0	10	5	6	0	0	0
2000-256-10	2	8	5	7	0	0	0
2000-1000-10	1	7	3	5	0	0	0

Table 10.4: Number of times, out of 10 execution, a known (numbered) coding mistake resulted in a detectable failure of Firefox running on a given hardware configuration (cpu speed-memory-disk size). Data from Syed, Robinson and Williams.<sup>132</sup>

An analysis of the statistical power of an experiment investigating the difference between proportions (i.e., the percentage of observed failures) needs to know the value of the proportions, the number of runs (10 in this case) and the desired p-value (0.05); to simplify things, Figure 10.22 uses the value of the lowest proportion and the difference between it and the higher proportion. The upper plot shows the power achieved (y-axis), if there does exist a given difference in proportions (x-axis), the proportions 0.05, 0.25 and 0.5 are plotted (the result is symmetric about 0.5 and so the lines for 0.75 and 0.95 would be the same as 0.25 and 0.05 respectively); where there were 10 and 50 runs involving the same fault case.

The probability of a difference being detected in results from 10 runs is well below 0.5 (i.e., less than a 50% chance of detecting a difference at a p-value of 0.05 or better).

The lower plot in Figure 10.22 shows the number of runs that need to be made to have an 80% chance of detecting, between two different hardware configurations, the difference in proportion listed on the x-axis, at a significance of 0.05.

If hardware characteristics account for only 10% of the difference in failure rate, over 100 runs would be needed to detect it.

The reason this experiment did not find a significant correlation between observed failure rate and hardware configuration is because of the small number of runs (i.e., 10) for each known fault.

Benchmarking to test impact of code-checking options...?

## 10.7 Meta-Analysis

Some issues are sufficiently interesting that they have been the subject of multiple studies. Meta-analysis is the process of systematically reviewing all the evidence available from multiple studies to produce a combined result.

Historically only descriptive statistics of samples has been available, rather than the raw data, and meta-analysis has primarily involved combining these values to produce a big picture result. When the raw data from more than one study is available, it may be possible to combine it into a form that can be used to repeat the analysis with a larger sample size.

The `meta` package ...

A study by Sabherwal, Jeyaraj and Chowa<sup>1188</sup> performed a meta-analysis to compute a correlation matrix based on 612 findings from 121 studies published between 1980 and 2004...

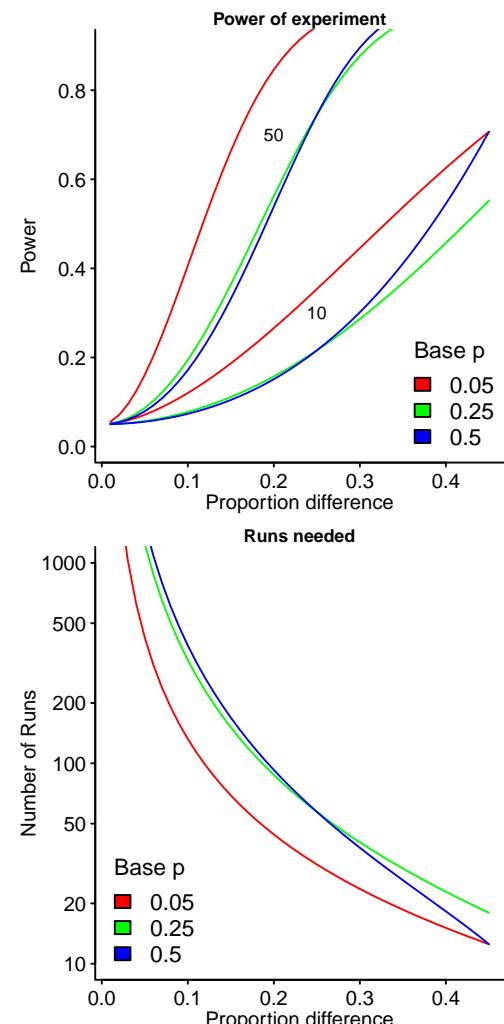


Figure 10.22: Power analysis (50 and 10 runs at various p-values) of detecting a difference between two runs having a binomial distribution (runs needed to achieve power=0.8 at various p-values). [code](#)



# Chapter 11

## Regression modeling

### 11.1 Introduction

Regression modeling is the default hammer used in this book to solve data analysis problems in software engineering.<sup>i</sup> The tree diagram, Figure 11.1, gives a high level overview of the various kinds of hammers available in the regression modeling toolkit. Concentrating on a single, general technique, removes the need for developers to remember how to select from, and use, many special case techniques (which in many cases only return a subset of the information returned from regression modeling).

The tree diagram connects the various regression techniques using the characteristics of the data they are designed to handle, with the techniques in red being the common use cases for their respective data characteristics.

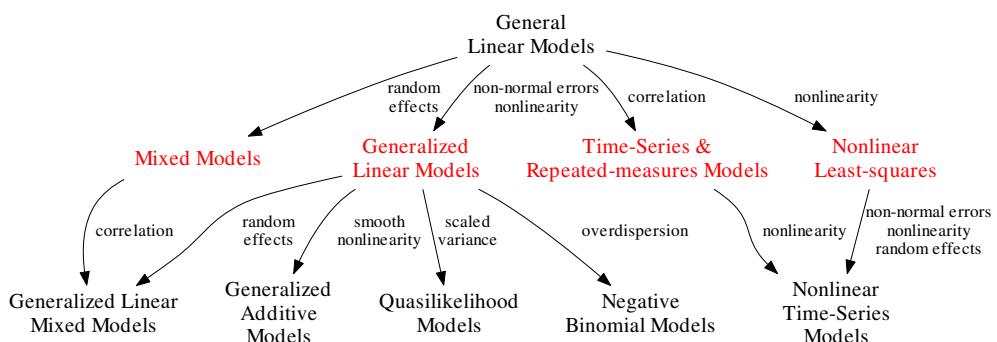


Figure 11.1: Relationship between data characteristics (edge labels) and applicable techniques (node labels) for building regression models.

Regression modeling is powerful enough to fit almost any data to any desired degree of accuracy, which means overfitting is an ever present danger and model validation (e.g., how well a model might fit new data or an estimation of the benefit obtained from including each coefficient in a model) is an important self-correcting step.

As always, it is necessary to remember the adage: “All models are wrong, but some are useful.”

The two main reasons for building a regression model are:

- understanding: by combining the explanatory variables into an equation that can be used to interpret why the response variable behaves the way it does,
- prediction: that is predicting the value of the response variable, for given values of the explanatory variables.

The focus of predictive modeling is accuracy of prediction and there is a willingness to trade-off understanding what is going on for greater accuracy, while the focus of interpretive modeling is understanding why and this creates a willingness to trade-off prediction accuracy for model simplicity.

<sup>i</sup> Machine learning is a not a regression modeling technique and is covered later.

Understanding is the primary focus for most of the model building in this book; builders of computing systems are generally interested in controlling what is happening, with predicting being a fall back position, and control requires understanding. Model building for prediction is often easier than building for understanding, so readers should not find it difficult switching to a predictive focus.

Regression models contain a *response variable*, one or more *explanatory variables*<sup>ii</sup> and some form of error term.

The *response variable* is modeled as some combination of *explanatory variables* and an additive or multiplicative error term (the error term associated with each explanatory variable represents behavior not accounted for by the explanatory variable; different kinds of regression model make different assumptions about the characteristics of the error term).

It is always possible to build a model that fits the data to within some degree of error, i.e., the amount of variation in the measurements that the model does not explain. It is very important to always ask how well a model fits the known data, not just the data used to build it.

If a sample contains many variables, then it is sometimes possible to build a model that has an impressive fit to the chosen response variable using only a few of the other variables. A study by Zeller, Zimmermann and Bird<sup>1484</sup> built a fault prediction model whose performance was comparable to the best available at the time. The model used four explanatory variables to predict the probability of a fault being reported in the source code contained in each file; the explanatory variables were the percentage occurrence of each of the letters IROP in the source code of each file. The model was *discovered* by checking how good a job every possible source code character did at predicting fault probability.

## 11.2 Linear regression

The simplest form of regression model is linear regression, where the *response variable* is modeled as a linear combination of *explanatory variables* and an additive error (the error terms are assumed to be independent and identically distributed;  $\varepsilon$  denotes the total error). The equation is:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

Note that the term *linear* refers to the coefficients of the model, i.e.,  $\beta$ , not the form taken by the explanatory variables which may have a non-linear form, as in:

$$y = \alpha + \beta x^2 + \varepsilon$$

or:

$$y = \alpha + \beta \log(x) + \varepsilon$$

A linear model is perhaps the most commonly used regression model, reasons for this include:

- many real world problems exhibit linear behavior, or a good enough approximation to it for practical purposes, over their input range,
- they are much easier to fit manually than more sophisticated models and until recently software to build other kinds of models was not widely available,
- they can generally be built with minimal input from the user (apart from having to decide which column of the data is the response variable).

The `glm` function<sup>iii</sup> builds a linear model and the use common case requires two argument value, a formula expressing a relationship between variables (response variable on the left and explanatory variable(s) on the right) and an object containing the data (this object is required to contain columns whose names match the identifiers appearing in the formula).

---

<sup>ii</sup> Books that concentrate on the predictive aspect of models use the term *prediction variable* or just *predictor*, while those that concentrate on running experiments use terms such as *control variables* or just the *controls*.

<sup>iii</sup> Many statistics books start by discussing the `lm` function, rather than `glm`, because the mathematics that underpins it is easier to learn; if you dear reader want to learn this mathematics I recommend taking this approach. As its name implies the Generalised Linear Method has a wider range of applicability and its use here is in line with the aim of teaching one technique that can be used everywhere. Also, the mathematics behind `glm` make fewer assumptions about the sample characteristics, e.g., it does not require the variance in the error to be constant (which `lm` does).

The formula has the form of an equation, with the  $=$  symbol replaced by  $\sim$  (pronounced *is distributed according to*) and the coefficients  $\alpha$  and  $\beta$  are implicitly present, i.e., they do not need to be explicitly specified.

The following code uses `glm` to build a model showing the relationship between the number of lines of source code (*sloc*) in FreeBSD and the number of days elapsed since the project started (in 1993):

```
BSD_mod=glm(sloc ~ Number_days, data=bsd_info)
```

The fitted equation is:

$$E[sloc] = \alpha + \beta \times Number\_days$$

where  $E[sloc]$  is the expected value of *sloc* (the error term is discussed below).

Figure 11.2 shows the measurement data points and the straight line whose coefficients are contained in the object returned by `glm`.

The `summary` function takes the object returned by `glm` and prints details about the fitted model;<sup>iv</sup> the following is for the model fitted to the FreeBSD data: `code`

Call:

```
glm(formula = sloc ~ Number_days, data = kind_bsd)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-82990	-32136	-3609	35389	87324

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )							
(Intercept)	1.139e+05	1.171e+03	97.24	<2e-16 ***							
Number_days	3.937e+02	4.205e-01	936.33	<2e-16 ***							
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

(Dispersion parameter for gaussian family taken to be 1657283104)

```
Null deviance: 1.4610e+15 on 4826 degrees of freedom
Residual deviance: 7.9964e+12 on 4825 degrees of freedom
AIC: 116172
```

Number of Fisher Scoring iterations: 2

The table following the **Coefficients:** header, in the `summary` output, lists estimated fitted values for  $\alpha$  and  $\beta$  (Intercept and *Number\_days* respectively), the standard error in these estimates (Std.Error) and (in the `Pr(>|t|)` column) the probability that if the true value of the coefficient was zero the estimated value would have occurred by chance.

The values listed in the `summary` output can be plugged into the model formula to give the following fitted equation:

$$sloc = 1.139 \cdot 10^5 + 3.937 \cdot 10^2 Number\_days$$

The fit between the model and the data is not perfect and the following are the two forms of uncertainty, or variation, present in the model:

1. Uncertainty in the values of the model coefficients. The values listed in the `Std. Error` column denote one standard deviation, which when added to the model gives the following:

$$sloc = (1.139 \cdot 10^5 \pm 1.171 \cdot 10^3) + (3.937 \cdot 10^2 \pm 4.205 \cdot 10^{-1}) Number\_days$$

2. Uncertainty caused by the inability of the explanatory variable used in the model to explain everything. This uncertainty is the  $\epsilon$  appearing in equation <??>; the term *residual* is used to denote this quantity. In the general case it is unlikely that  $\epsilon$  will have a fixed value over the range of values supported by a model and `glm` does not generate a single value.

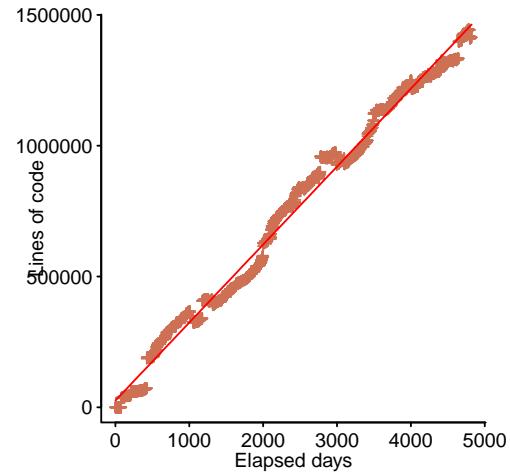


Figure 11.2: Total lines of source code in FreeBSD by days elapsed since the project started (in 1993). Data from Herraiz.<sup>1333</sup> [code](#)

<sup>iv</sup> Only a few digits of the estimated values are printed by default.

In Figure 11.2 the variations in the unexplained error,  $\epsilon$ , appear to be small. Assuming a fixed value, a call to `aov` returns 40,710 as the residual standard error. The equation including this estimate of the residual is:

$$sloc = 1.139 \cdot 10^5 + 3.937 \cdot 10^2 Number\_days \pm 4.071 \cdot 10^4$$

In other words the difference between measured values and values predicted by this fitted model will have a standard error of  $4.071 \cdot 10^4$ .

The object returned by the call to `glm` can be used to make predictions and these can be overlaid on the output from an earlier call to `plot`, as follows:

```
BSD_pred=predict(BSD_mod) # predict using measured values
lines(BSD_pred, col="red") # x-axis start at 1 and increment
```

The `predict/lines` approach follows this book's aim of using techniques that work for the general case. Plotting a fitted straight line is such a common operation that there is a function for doing just that, e.g., `abline(reg=BSD.enalty5000_mod, col="red")`, but this does not always work when the axis have been scaled in some way and is of no use for fitted models that are more complicated than a straight line.

Before being carried away with the high degree of agreement between this model and the data it is important to remember that the model has a number of characteristics that do not reflect reality, including:

- source code does not spontaneously grow of its own accord and the only justification for treating *number of days* as an explanatory variable is that the resulting model provides potentially interesting insight into the rate of growth of these software systems.
- when it started the BSD project contained zero lines of code, but this model has an Intercept of  $1.39 \cdot 10^5$ ,
- the model shows the number of lines increasing for ever, at a constant rate, whereas at some point in the future growth must slow down and eventually stop,
- it says nothing about large amounts of code being added/removed over very short periods (known to exist because of the visible breaks in the connectedness of the plotted values).

While the model has various disconnects with reality, it does provide strong evidence that growth has been remarkable constant over a long period. Unless there are seismic changes within the FreeBSD development world the constant rate of code growth would be expected to continue to hold for a non-trivial number of days into the future.

The call to `summary`, passing the value returned by `glm`, is an example of function overloading in action. The value returned by `glm` has class `glm`, which when passed as an argument to `summary` results in `summary.glm` being called; a call to `predict` results in `predict.glm` being called (function overloading is the most common use of object oriented constructs in R programs; the use of a period in the function name is a naming convention followed by the implementers and not something automatically added by the compiler).

Some of the factors and processes that might be involved in driving a fixed rate of growth over 20 years include:

- developers working on the system have continually found new functionality to add,
  - if there has always been functionality to add, why haven't more developers become involved to increase the rate of growth until there is less to do?
  - to what extent is the continual stream of new hardware devices responsible for driving growth?
- what are the bottlenecks that have prevented increases in growth rate when the resources have been available?
  - has growth rate remained constant because the developers working on the systems have remained constant?
  - is there a buffer of code waiting to be released, whose growing and shrinking hides an internal growth rate that is much more variable than the externally visible rate?

### 11.2.1 Scattered measurement values

In the previous example the measurements ran together in a way that created a visually recognizable line. The common case is not always so accommodating and when many samples are plotted a scattering of disjoint points often appears; viewed as a whole a general trend may emerge.

A study by Kampstra and Verhoef<sup>731</sup> investigated the estimated cost and duration of 73 large Dutch federal IT projects.<sup>v</sup> Figure 11.3 shows that very few of the measurement points are on the (red) line specified by the model returned by `glm`; the variability of the measured values is much larger than that for the FreeBSD model. While numeric estimates of the uncertainty present in the fitted model are readily available, interpreting these numeric values requires a degree of effort and some experience. A confidence interval provides an easy to interpret visual representation of the uncertainty in a fitted model.

The kind of uncertainty of interest will depend on whether the model is built to gain understanding or make predictions:

- when understanding is the priority, the confidence interval of interest involves the estimated model coefficients:

- a call to `predict` with the `se.fit=TRUE` argument returns the standard error for each fitted value. Multiplying `se.fit` by `qnorm`<sup>vi</sup> converts the returned value to a 95% confidence interval (2.5% above and below the fit; the two `qnorm` values differ only in sign because the Normal distribution is symmetrical), i.e., there is a 95% expectation that the actual model fits within the interval enclosed by these lower/upper bounds. `qnorm(0.975)==1.96` and the literal value is often used (in fact the value 2 is often seen as a sufficiently close approximation).<sup>vii</sup>

```
fed_pred=predict(fed_mod, newdata=list(log.IT=1:7, log.IT_sqr=(1:7)^2),
                 se.fit=TRUE)
lines(fed_pred$fit, col="green")      # fitted line
# CI above and below
lines(fed_pred$fit+qnorm(0.975)*fed_pred$se.fit, col="green")
lines(fed_pred$fit+qnorm(0.025)*fed_pred$se.fit, col="green")
```

- the `confint` function in the `MASS` package or the `boot.ci` function in the `boot` package can be used to obtain a point estimate of the confidence interval of the fitted model coefficients.
- when prediction is the priority the interval is known as the *prediction interval*; the bounds between which newly measured values are expected to appear. Two sources of uncertainty are added to calculate the prediction interval: uncertainty in the model coefficients (i.e., the confidence interval) plus the variance in the data not explained by the fitted model.

```
# print.aov also calculates it from residuals returned by glm...
MSE=sum(fed_mod$residuals^2)/(length(fed_mod$residuals)-2)
# Variances, but not sd, can be added
pred_se=sqrt(fed_pred$se.fit^2+MSE)
lines(fed_pred$fit+1.96*pred_se, col="blue")
lines(fed_pred$fit-1.96*pred_se, col="blue")
```

When measurement values and a fitted regression line are plotted, it is easy to visually fixate on the line and forget about the associated uncertainties. Including a confidence band as part of a plot provides a vivid visual reminder of the quality of the fit.

Plotting values does not always reveal an obvious pattern in the points. A visual pattern may not exist because no relationship exists between the response and explanatory variables or because the noise in the data is much greater than the signal (i.e., a relationship that does exist is swamped by the noise present in the measurements).

How much random scattering of measurement values has to exist before a fitted regression model can be said to be not worth bothering about?

<sup>v</sup> They discovered that there was a lot of uncertainty in the estimates given.<sup>?</sup>

<sup>vi</sup> This calculation assumes that the measurement error has a Normal distribution, the default assumption made by `glm` when building a model.

<sup>vii</sup> For small sample sizes a call to `qt` may be more accurate.

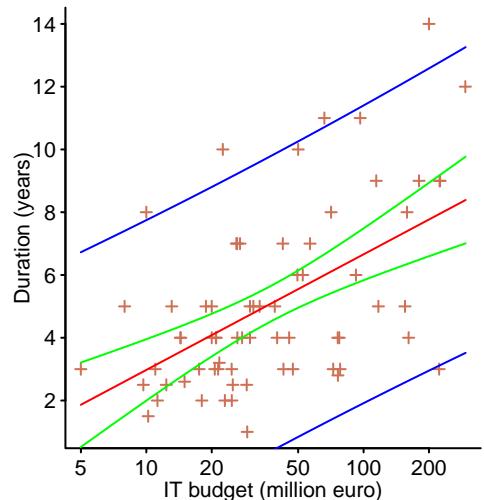


Figure 11.3: Estimated cost and duration of 73 large Dutch federal IT projects, along with fitted model and 95% confidence intervals. Data from Kampstra et al.<sup>731</sup> [code](#)

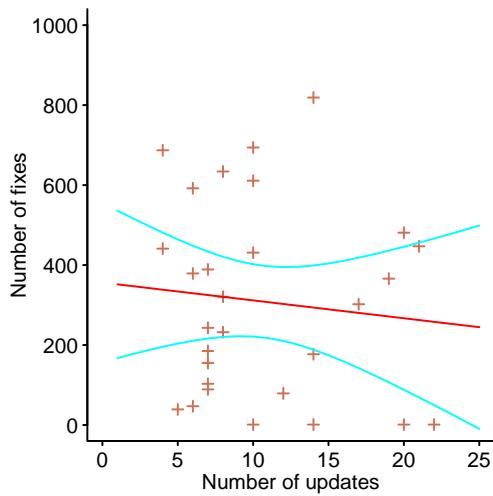


Figure 11.4: Number of updates and fixes in each Linux release between version 2.6.11 and 3.2. Data from Corbet et al.<sup>288</sup> [code](#)

The `glm` function and many other model building functions available in R are capable of fitting models to data points that are randomly distributed. For instance, Figure 11.4 shows the number of updates and fixes made in various Linux versions released between early 2011 and 2012. The standard error of the fitted line show that its slope could have a positive or negative value.

The output from `summary` shows how poor the fit actually is; the `Pr(>|t|)` column lists the p-value for the hypothesis that the coefficient in the corresponding row is zero, i.e., that no relationship was found to exist for that component of the model. [code](#)

Call:  
`glm(formula = Fixes ~ Total.Updates, data = cleaned)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-310.60	-223.67	0.48	184.51	525.26

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	356.233	101.522	3.509	0.0016 **
Total.Updates	-4.464	8.478	-0.526	0.6029

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 60685.71)

Null deviance: 1655335 on 28 degrees of freedom  
 Residual deviance: 1638514 on 27 degrees of freedom  
 AIC: 405.62

Number of Fisher Scoring iterations: 2

## 11.2.2 Discrete measurement values

Regression models are not limited to using continuous numeric explanatory variables, variables having nominal values can also be used.

A study by Cook and Zilles<sup>284</sup> investigated the impact of compiler optimization flags on the ability of software to continue to operate correctly when subject to random bit-flips, i.e., simulating random hardware errors; 100 evenly distributed points in the program were chosen and 100 instructions from each of those points were used as fault injection points, giving a total of 10,000 individual tests run, for each of 12 programs from the SPEC2000 integer benchmark compiled using gcc version 4.0.2 (using optimization options: 00, 02 and 03) and the DEC C compiler (called *osf*).

The fitted model has percentage of correct benchmark program execution as the response variable and optimization level as the explanatory variable; the following is the call to `glm`:

`bitflip_mod=glm(pass.masked ~ opt_level, data=bitflip)`

The following is `summary` output of the fitted model: [code](#)

Call:  
`glm(formula = pass.masked ~ opt_level, data = bitflip)`

Deviance Residuals:

Min	1Q	Median	3Q	Max
-12.6689	-2.8454	-0.3478	4.4017	8.1100

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	28.589	1.825	15.665	< 2e-16 ***
opt_level02	9.161	2.581	3.550	0.00112 **
opt_level03	7.429	2.581	2.878	0.00677 **

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 29.97578)

Null deviance: 1785.8 on 38 degrees of freedom  
 Residual deviance: 1049.2 on 35 degrees of freedom  
 AIC: 249.07

Number of Fisher Scoring iterations: 2

Plugging the model coefficients into the regression equation we get:

$$\text{pass}.\text{masked} = 28.6 + 9.2 \cdot D_{O2} + 7.4 \cdot D_{O3} + 11.6 \cdot D_{osf}$$

where:  $D_i$ , known as a *dummy variable* or *indicator variable*, take one of two values:

$$D = \begin{cases} 1 & \text{optimization flag used} \\ 0 & \text{optimization flag not used} \end{cases}$$

The value for optimization 00 is implicit in the equation, it occurs when all other optimizations are not specified, i.e., its value is that of the intercept.

The standard error in the O2 and O3 compiler options is sufficiently large for their respective confidence bounds to significantly overlap. This suggests that these two options have a similar impact on the behavior of the response variable.

### 11.2.3 Uncertainty only exists in the response variable

The algorithms used to fit regression models often attempt to minimise the difference between the measured points and a specified equation. For instance, least-squares minimises the sum of the squares of the distance along one axis between each data point and the fitted equation;<sup>viii</sup> alternative minimization criteria are discussed later, e.g., giving greater weight to positive error than negative error.

An important, and often overlooked, detail, is that many regression techniques assume that the values of the explanatory variable(s) contain no uncertainty (e.g., measurements are exact), with all uncertainty,  $\varepsilon$ , occurring in the response variable (see the [first equation](#) at the start of this chapter).

A consequence of assuming uncertainty only exists in the response variable, is that the equation produced by fitting a model that specifies, say,  $X$  as the explanatory variable and  $Y$  the response variable will not be consistent with a model that assumes  $Y$  is the explanatory variable and  $X$  the response variable. That is, algebraically transforming the first equation produces an equation whose coefficients are different from the second.

Take as an example, data from Kroah-Hartman<sup>548</sup> who measured the number of commits made between the release of a version of Linux and the immediately previous version, and the number of developers who contributed code to that release, for the 67 major kernel releases between versions 2.6.0 and 4.6.

In the upper plot of Figure 11.5, the number of developers is treated as the explanatory variable (x-axis) and number of commits as the response variable (y-axis), with the fitted regression line in red and dashed lines showing the difference between measurement and fitted model. In the lower plot the explanatory/response roles played by the two variables when fitting a regression model is switched; to simplify comparison the axis denote the same variables in both plots, with the green line denoting the fitted model and dashed lines showing the difference between measurement and model (now on the x-axis response variable; the line fitted in the upper plot is also given for comparison, still in red).

In the first case the fitted equation is:

$$\text{commits} = -237 \pm 523 + (8.7 \pm 0.44) \cdot \text{Number\_devs}$$

transforming this equation we get:

$$\begin{aligned} \text{Number\_devs} &= \frac{237 + \text{commits}}{8.7} \\ &= 27 + 0.11\text{commits} \end{aligned}$$

<sup>viii</sup> Minimising the sum of squares in the error has historically been popular because it is a case that can be analysed analytically.

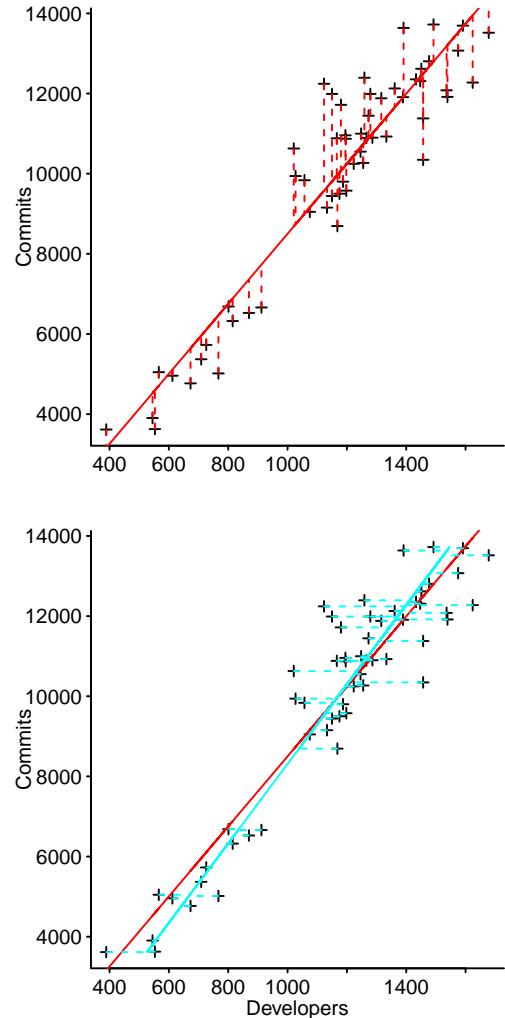


Figure 11.5: The number of commits made and the number of contributing developers for Linux versions 2.6.0 to 3.12. The green line in the right plot is the regression model fitted by switching the x/y values. Data from Kroah-Hartman.<sup>548</sup> [code](#)

However, if a new model is fitted by switching the roles of the two variables in the formula passed to `glm`, the returned model is described by the following equation:

$$\text{Number\_devs} = 162 \pm 52 + (0.10 \pm 0.005) \cdot \text{commits}$$

which is different from the equation obtained by transforming the first fitted model.

There is another difference between the two fitted models, the second model has a better fit to the data. Somebody only interested in the quality of fit might be tempted to select the second model, purely for this reason.

What is the procedure for deciding which measurement variables play the role of response and explanatory variable, e.g., should number of developers be considered an explanatory or response variable?

An important attribute of explanatory variable(s) is that their value is controlled by the person making the measurement. For instance, the model building process used to create Figure 11.2 had number of days as the explanatory variable; this choice was completely controlled by the person making the measurements.

The Kroah-Hartman commit measurements are based on the day of release of a version of the Linux kernel, a date that is outside the control of the measurement process. In fact both measurements have the characteristics of a response variables, that is, the value they have, was not selected by the person making the measurement. The possibility of variation in Linux version release dates is a source of uncertainty that needs to be treated as a form of measurement error.

Building regression models using explanatory variables containing measurement error can result in models containing biased and inconsistent values, as well as inflating the Type I error rate.<sup>192, 1236</sup>

There are a variety of regression model building techniques that take into account error in the explanatory variable. These techniques are sometimes known as *model II* linear regression techniques (model I being the case where there is no uncertainty in the explanatory variables), *errors-in-variable models*, *total least-squares* or *latent variable models*; using methods such as *major axis*, *standard major axis* and *ranged major axis*.

If all the variables used to build a model contain some amount of error, then it is necessary to decide how much error each variable contributes to the total error in the model. Some model II techniques are not scale invariant, that is they are only applicable if both axis are dimensionless or denote the same units, otherwise rescaling one axis (e.g., converting from kilometers to miles) will change its relative contribution; if each axis denotes a different unit it does not make sense to use a model building technique that attempts to minimise some measure of combined uncertainty.

SIMEX (SIMulation-EXtrapolation) is a technique for handling uncertainty in explanatory variables that works in conjunction with a range of regression modeling techniques. While the SIMEX approach does not suffer from many of the theoretical problems that other techniques suffer from, it does require that the model builder provide an estimate of the likely error in the explanatory variable. The `simex` package implements this functionality and supports a wide variety of regression models built by functions from various packages.

Continuing with the Linux developer/commit count example, to build a regression model using SIMEX we need an estimate of the uncertainty in the number of developers contributing at least one commit to any given release. The `simex` function taking a model built using `glm` (and by other regression model building functions) and an estimate of the uncertainty in one or more of the explanatory variables and returns an updated model that takes this uncertainty into account.

The following is a rough and ready approach to estimating the uncertainty in the Kernel attributes measured by Kroah-Hartman:

- the release date of a new version of Linux is assumed to have an uncertainty of  $\pm 14$  days about the actual release date.<sup>ix</sup>
- the possible variation in the unique contributor count for any release is assumed to be uniformly distributed in the range: measured contributor count plus/minus number of developers contributing their first commit in the last 14 days.

---

<sup>ix</sup> Pointers to a more reliable, empirically derived, value are welcome.

- making the above assumptions we get a standard deviation of 41 for the number of unique developers making at least one commit, averaged over all versions (see [regression/clean/dev-commit.R](#)).

This estimate of the standard deviation in the explanatory variable is integrated into a regression model that takes account of uncertainty in more than just the response variable as follows:

- first build a regression model using `glm` in the usual way, but with the optional named parameter `x` set to TRUE (`y` also needs to be TRUE, but this is its default value and so the assignment below is redundant),
- pass the model returned by `glm` to `simex`, along with the name of the explanatory variable and its estimated standard deviation.

```
yx_line = glm(commits ~ developers, x=TRUE, y=TRUE)

sim_mod=simex(yx_line, SIMEXvariable="developers", measurement.error=41)
```

The `summary` output (see [regression/dc-simex.R](#)) shows that the model returned by `simex` is the following:<sup>x</sup>

$$\text{commits} = -387 \pm 453 + (8.9 \pm 0.4) \cdot \text{Number\_devs}$$

The error in each individual explanatory variable measurement can be specified by assigning a vector to `measurement.error` (the argument `asymptotic=FALSE` is also required); see Figure 11.35.

There are techniques, and R packages, for building complete models starting from the data, rather than refitting an existing regression model. For those wanting to a build model from scratch the `lmodel2` package provides functions that implement many of the available methods.

How reliable is a fitted model that has been built by ignoring any uncertainty/error in explanatory variable measurements? The only way to answer this question is to build a model that takes this error into account and compare it with one that does not.

It is wrong to assume that a model fitted by switching response/explanatory variables will fall within the 95% confidence intervals of the existing model, as Figure 11.6 shows. One Effort/Size fit has a slope greater than one and the other less than one, with the value one being an important dividing line in the interpretation of behavior (i.e., do economies of scale exist for software development).

Many of the measurement values treated as explanatory variables in this book were not under the control of the person who measured them. For instance, lines of code, number of files and reported problems measured at a given point in time are all response variables. To reduce your authors workload most of the model fitting in this book does not make any adjustments for errors in the explanatory variables.

#### 11.2.4 Modeling data that curves

A model based on a straight line is a wonderful thing to behold, it is simple to explain and often aligns with people's expectations (many useful real world problems are well fitted by a straight line). However, life is complicated and throws curved data at us.

Having encountered an operating system having constant lines of code growth over many years, it is tempting to draw a conclusion about the growth rate of other operating systems. However, the way in which the data points curve around the fitted line in the upper plot of Figure 11.7 suggests that some of the processes driving the growth of the Linux kernel are different from those driving FreeBSD; perhaps a quadratic or exponential equation would be a better fit (these possibilities were chosen because they are two commonly occurring forms for upwardly curving data).

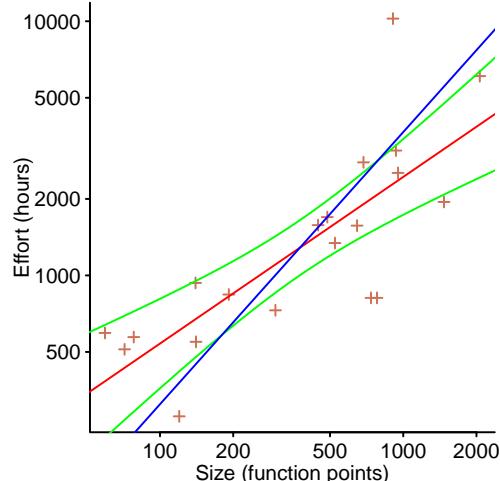


Figure 11.6: Effort/Size of various projects and regression lines fitted using Effort as the response variable (red, with green 95% confidence intervals) and Size as the response variable (blue). Data from Jørgensen et al.<sup>716</sup> code

<sup>x</sup> Readers might like to experiment with the value of the `measurement.error` to see the impact on the model coefficients.

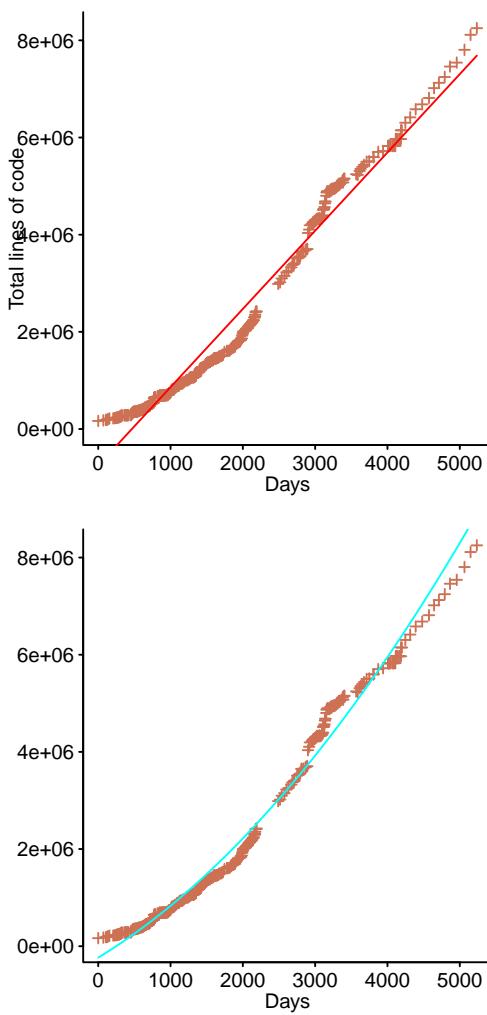


Figure 11.7: Lines of code in every initial release (i.e., excluding bug-fix versions of a release) of the Linux kernel since version 1.0, along with fitted straight line (upper) and quadratic (lower) regression models. Data from Israeli et al.<sup>673</sup> [code](#)

This section is about fitting linear models, so the possibility of an exponential fit is put to one side for the time being; building non-linear models, including better fitting non-linear models to this data, is discussed later.

The following call to `glm` fits an equation that is quadratic in the variable `Number_days`; the righthand side of the formula contains `Number_days+I(Number_days^2)`. The `I` (sometimes known as *as-is*) causes its argument to remain unevaluated and is treated as a distinct explanatory variable (without the `I`, `Number_days` would be squared, added to the first instance and the sum treated as a single explanatory variable). An alternative way of including a squared explanatory variable in the model is to assign the value `Number_days^2` to a new variable and include this new variable's name on the righthand side of the formula. The result of fitting this equation is shown on the lower plot of Figure 11.7.

```
linux_mod=glm(sloc ~ Number_days+I(Number_days^2), data=linux_info)
```

The quadratic fit looks like it could be better than linear, but perhaps a cubic, quartic or higher degree polynomial would be even better. The higher the order of the polynomial used, the smaller the error between the fitted model and the data used. The error decreases because the additional terms are used to adjust the model to do a better job of following the random fluctuations in the data. An Occam's razor method is needed to select the number of terms that produces the simplest model consistent with the data and having an acceptable error.

The Akaike Information Criterion, AIC, is a commonly used metric for comparing two or more models (available in the `AIC` function). It takes into account both how well a model fits the data and the number of free coefficients in the model; free coefficients have to pay their way by providing an appropriate improvement in a model's fit to the data.<sup>xi</sup> AIC can also be viewed as the information loss when the true model is not among those being considered.<sup>202</sup>

One set of selection criteria<sup>202</sup> are that models whose AIC differs by less than 2 are more or less equivalent, those that differ by between 4 and 7 are clearly distinguishable, while those differing by more than 10 are definitely different.

How much better does a quadratic equation fit Linux SLOC growth compared to a straight line and how much better do higher degree polynomials fit? The following lists the AIC for models fitted using polynomials of degree 1 to 4 (lower values of AIC are better). After initially decreasing the AIC starts to increase once a fourth degree polynomial is reached; the third degree polynomial is thus the better fitting linear polynomial of those tested, i.e., other forms of equation could be better. [code](#)

```
[1] Degree 1, AIC= 13998.0004739753
[1] Degree 2, AIC= 13674.6883243397
[1] Degree 3, AIC= 13220.8542892188
[1] Degree 4, AIC= 13221.7072389496
```

The following is the `summary` output of the fitted cubic model: [code](#)

```
Call:
glm(formula = LOC ~ Number_days + I(Number_days^2) + I(Number_days^3),
     data = latest_version)

Deviance Residuals:
    Min      1Q   Median      3Q      Max 
-428217 -80061    6503   64889   620500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.432e+05  2.876e+04 11.935 < 2e-16 ***
Number_days -3.664e+02  5.144e+01 -7.123 3.79e-12 ***
I(Number_days^2) 8.167e-01  2.456e-02 33.258 < 2e-16 ***
I(Number_days^3) -9.184e-05  3.371e-06 -27.242 < 2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 23001633959)
```

<sup>xi</sup> A negative value may be the result of a nominal explanatory variable having many values, e.g., dates are represented as strings that are best converted using `as.Date`.

```
Null deviance: 1.8867e+15 on 494 degrees of freedom
Residual deviance: 1.1294e+13 on 491 degrees of freedom
AIC: 13221
```

Number of Fisher Scoring iterations: 2

Regression modeling finds the best fit for an equation over the range of the data used and AIC helps prevent overfitting. No claims are made about how well the model is likely to fit data outside the range that was used to build it. Using a model optimized to fit the available data to make predictions outside the interval of the data used can produce very surprising results.

What is the behavior of the cubic model outside its fitted range and in particular what predictions does it make about future growth? This model predicts a future decrease in the number of lines of code (see Figure 11.8). A decreasing number of lines is the opposite of previous behavior this prediction is unlikely to be believed by many people (if this behavior were predicted by a more detailed model of code growth that closely mimicked real-world development by using information on the number of developers actively involved and a list of functionality that is likely to be implemented, it might be more believable).

A quadratic equation might not fit the data as well as a cubic equation, but the form of its predictions (increasing growth) is consistent with expectations.

If the purpose of the model is understanding, then the quadratic model maps more closely to anticipated behavior; if the purpose is prediction within the interval of the fitted data, then the cubic model is likely to have a smaller error.

What about fitting other kinds of equations to the data? Equations such as  $Y = \alpha e^{\beta X} + \epsilon$  and  $Y = \alpha X^\beta + \epsilon$  are nonlinear in  $\beta$ ; non-linear model building is covered later.

For a software system to grow more code has to be added to it than is deleted. A constant rate of growth suggests either a constant amount of developer effort or a bottleneck holding things up; an increasing rate of growth (i.e., quadratic) suggests an increasing rate of effort. The different code growth pattern seen in the Linux kernel, compared to NetBSD/FreeBSD and various other applications, has been tracked down to device driver development;<sup>506</sup> new hardware devices often share many similarities with existing devices and for Linux developers tend to copy an existing driver, modifying it to handle the hardware differences; it is this reuse of existing code that is the source of what appears to be a non-linear growth in developer effort.

This method of creating a new device driver, performed by many developers working independently, can continue for as long as there are new devices coming to market; the evolution of Linux device drivers is discussed elsewhere...

A linear regression model is not limited to using polynomials of explanatory variables, any function can be used as long as the coefficients of the model occur in linear form. For instance, the FreeBSD model plotted in Figure 11.2 might include a seasonal term that varies with time of year; while a model containing the term  $A \sin(2\pi ft + \phi)$  is nonlinear (because of  $\phi$ , the phase shift<sup>xii</sup>) it can be written in the following linear form:

$$A \sin(2\pi ft + \phi) = \alpha_s \sin(2\pi ft) + \alpha_c \cos(2\pi ft)$$

where:  $\alpha_s = A \cos \phi$  and  $\alpha_c = A \sin \phi$ ;  $A = \sqrt{\alpha_s^2 + \alpha_c^2}$  and  $\phi = \arctan \frac{\alpha_c}{\alpha_s}$ .

The call to `glm` is now (the argument to the trig functions has to be expressed in radians):

```
rad_per_day=(2*pi)/365
freebsd$rad_Number_days=rad_per_day*freebsd$Number_days
season_mod=glm(sloc ~ Number_days+
  I(sin(rad_Number_days))+I(cos(rad_Number_days)),
  data=freebsd)
```

The summary output from the fitted model shows that while a seasonal component probably exists, its overall contribution is very small (see [regression/Herraiz-BSD-season.R](#)).

While fitting a model using all available measurements points is a reasonable first step, subsequent analysis may suggest that the sample might best be treated as two or more disjoint samples. There may be time dependent factors that have a strong influence on growth patterns.

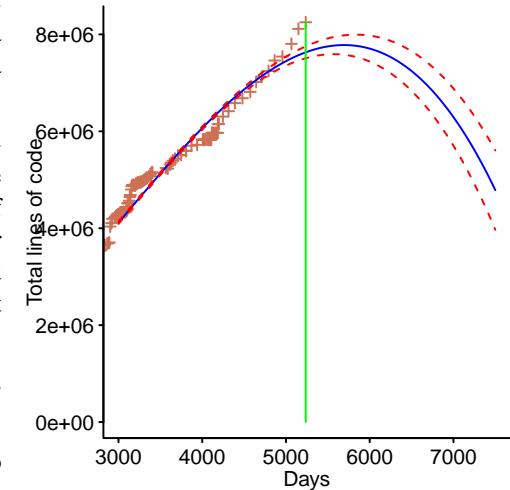


Figure 11.8: Actual (left of vertical line) and predicted (right of vertical line) total lines of code in Linux at a given number of days since the release of version 1.0, derived from a regression model built from fitting a cubic polynomial to the data (dashed lines are 95% confidence bounds). Data from Israeli et al.<sup>673</sup> [code](#)

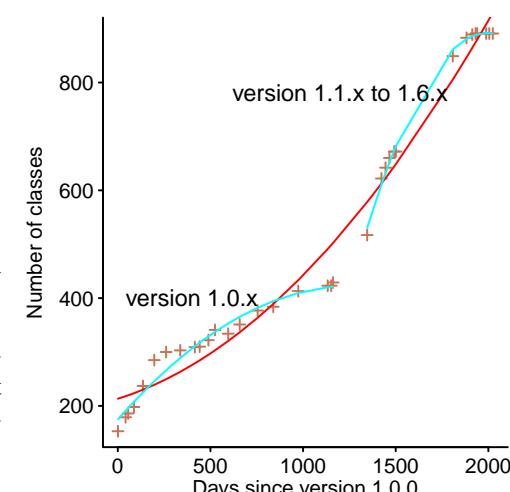


Figure 11.9: Number of classes in the Groovy compiler at each release, 2013 days since version 1.0. Data From Vasa.<sup>1394</sup> [code](#)

<sup>xii</sup>Some texts use  $A \cos(2\pi ft + \phi)$ , which changes the phase by 90° and changes some signs

Figure 11.9 shows the number of classes in the Groovy compiler at each release, in days since version 1.0. There are noticeable kinks in the growth rate at around 1,300 and 1,500 days. Fitting a model to the complete sample shows upward trending quadratic growth in the number of classes over time, while fitting separate models to two halves of the sample shows quadratic growth that flattens out.

Some investigation finds that the kink occurs at a transition between version numbers. It is easy to invent a variety of explanations for the pattern of behavior seen, but treating the measurements as-if they came from a single continuously developed code base is probably not one of them; further investigation of the circumstances behind the development of the Groovy compiler is needed if a reasonable level of confidence is required in whatever model is finally selected.

## 11.2.5 Visualizing the general trend

Even when the measurement points are scattered in what appears to be an obvious general direction, it is worthwhile quickly obtaining an estimate of the trend followed by the data.

A general technique for highlighting a trend follows by data is to fit a regression model to a consecutive sequence of small intervals of the data and join this sequence of fits together to form a continuous line. Two methods based on this idea (both fitting so the lines smoothly run together) are LOWESS (LOcally WEighted Scatterplot Smoothing) and LOESS (LOcal regrESSion); `lowess` and `loess` are the respective functions, with `loess` being used in this book.

A study by Kunst<sup>795</sup> counted, for 148 languages, the number of lines committed to Github (between February 2013 and July 2014) and the number of questions tagged with that language name on Stackoverflow.

The upper plot of Figure 11.10 shows lots of points which look as-if they trend in a straight line. The `loess` fit, red line in lower plot, shows the trend having a distinct curve. Experimenting with a quadratic equation in `log(lines_committed)` shows, blue line in lower plot, that this more closely follows the `loess` fit than a straight line (a quadratic fit has a lower AIC than a linear one; see [regression/langpop-corger-nl.R](#)).

A call to `loess` has the same pattern as a call to `glm`, with the possible addition of an extra argument; `span` is used to control the degree of smoothing:

```
loess_mod=loess(log(stackoverflow) ~ log_github, data=langpop, span=0.3)
x_points=1:max(langpop$log_github)
loess_pred=predict(loess_mod, newdata=data.frame(log_github=x_points))
lines(exp(x_points), exp(loess_pred), col=pal_col[1])
```

A study by Edmundson, Holtkamp, Rivera, Finifter, Mettler and Wagner<sup>375</sup> investigated the effectiveness of web security code reviews and asked professional developers to locate vulnerabilities in code.

The `lowess` fit, blue line in Figure 11.11, suggests that the percentage of vulnerabilities found increases as the number of years working in security increases, but then rapidly decreases; this performance profile seems unrealistic. A fitted straight line, in red, shows a decreasing percentage with years of work in the security field (its p-value is 0.02).

Perhaps the correct interpretation of this sample is that average performance does increase with years of working in the field, but that the subjects with many years working in security, who took part in the study, were more managerial and customer oriented people who had time available to take part in the experiment, i.e., this is a subject sampling problem. At the time of the study software security work was rapidly expanding, so the experience profile will be skewed with more subjects being less experienced.

When neither argument has been transformed, the value returned by the `loess.smooth` function can be passed directly to `lines`.

```
lines(loess.smooth(dev$experience, dev$written, span=0.5), col=pal_col[2])
```

A `loess` visualization can also be helpful when the number of data points is so large they coalesce into formless blobs. The Ultimate Debian Database, see Figure 11.23, is an example.

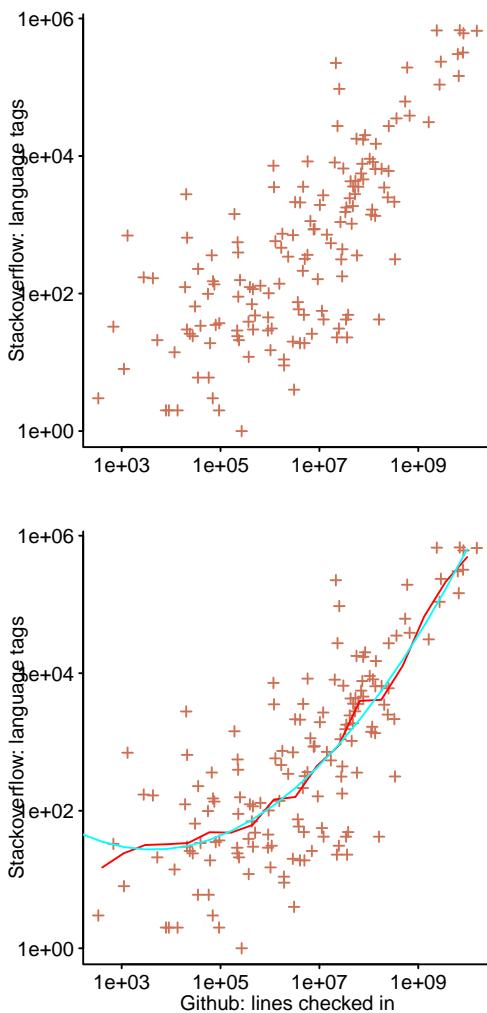


Figure 11.10: For each distinct language, the number of lines committed on Github and the number of questions tagged with that language. Data from Kunst.<sup>795</sup> [code](#)

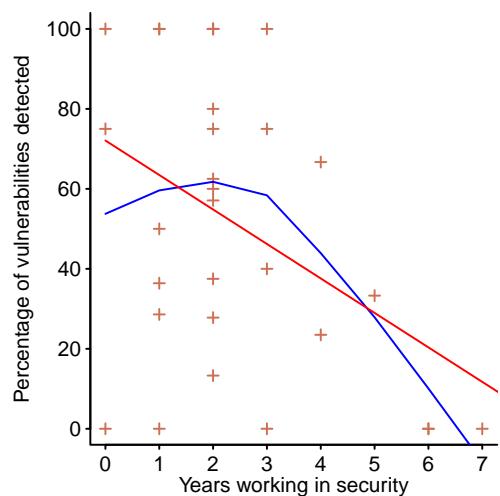


Figure 11.11: Percentage of vulnerabilities detected by developers working a given number of years in security. Data extracted from Edmundson et al.<sup>375</sup> [code](#)

The default behavior of the loess implementation is to divide the range of x-axis values into fixed intervals. When the range of x-values varies by many orders of magnitude the fitted curve can look over stretched at the low values and compressed at high values.

One solution is to reduce the range of x-values by, for instance, taking the log, smoothing and then expanding (see [regression/java-api-size.R](#)); the following code is used later:

```
t=loess.smooth(log(API$Size), API$APIs, span=0.3)
lines(exp(t$x), t$y, col=loess_col)
```

## 11.2.6 Influential observations and Outliers

Influential observations are observations that have a disproportionate impact on the values of the model coefficients, e.g., a single observation significantly changing the slope of the fitted straight line. The terms *leverage* or *hat-value* describes the amount of influence a data point has on a fitted model; the `hatvalues` function takes the model returned by `glm` and returns the leverage of each point.

Influential observations might be removed or modified, or regression techniques used that reduce the weight given to what are otherwise overly influential points (e.g., the `glmrob` function in the `robustbase` package).

Outliers are discussed as a general issue in Chapter 15, this subsection discusses outliers in the context of regression modeling. In the context of regression modeling an outlier might be defined as a data point having a disproportionately large standardized residual (here Studentized residuals are used).

To repeat an important point made in Chapter 15: excluding any influential observations or outliers from the analysis is an important decision that needs to be documented in the results.

*Cook's distance* (also known as *Cook's D*) combines leverage and outlierness into a single number that is a commonly used metric.

A study by Fenton, Neil, Marsh, Hearty, Radliński and Krause<sup>424</sup> involved data from 31 software systems for embedded consumer products. Figure 11.12 shows development effort against the number of lines of code, along with a fitted straight line and standard error bounds. At the right edge of the plot are two projects that consumed over 50,000 hours of effort and the number of lines of code for these projects looks very small in comparison with other projects. Is the fitted model overly influenced by these two projects and should they be ignored or adjusted in some way?

As the number of points in a sample grows there is an increasing probability that one or more of them will be some distance away from the fitted line; in any large sample a few apparent outliers are to be expected as a natural consequence of the distribution of the error. The following example illustrates the dangers of not taking sample size into account when making judgements about the outlier status of a measurement.

Figure 11.13 shows the results of building a model and removing measurements having both a high Cook's distance and Studentized residuals, and the repeating the process until points stop being removed. At the end of the process most of the measurements have been removed.

Removing overly influential points until everything looks respectable is seductive, it is an easy to follow process that does not require much thought about the story that the data might have to tell. For those who don't want to think about their data, the `outlierTest` function in the `car` package can be used to automate outlier detection and removal (it takes a model returned by `glm` and returns the Studentized residuals of points whose Bonferroni corrected p-value is below a cutoff threshold; default `cutoff=0.05`).

A method of visualizing the important influential observation and outlier information is required. The `influenceIndexPlot` function in the `car` package takes the model returned by `glm` and plots the Cook's distance, Studentized residual, Bonferroni corrected p-value and hat-value for each data-point; Figure 11.15 is for the Fenton et al data.

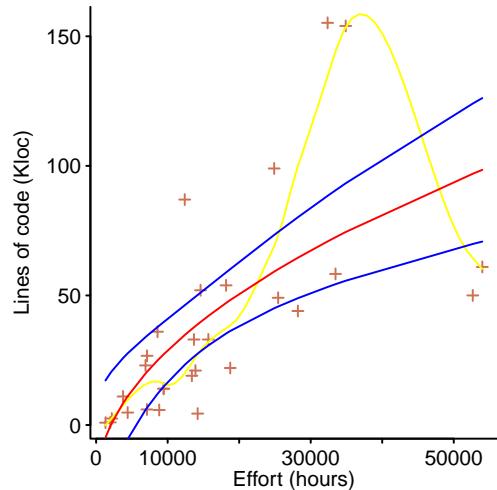


Figure 11.12: Hours to develop software for 29 embedded consumer products and the amount of code they contain, with fitted regression model and loess fit (yellow). Data from Fenton et al.<sup>424</sup> [code](#)

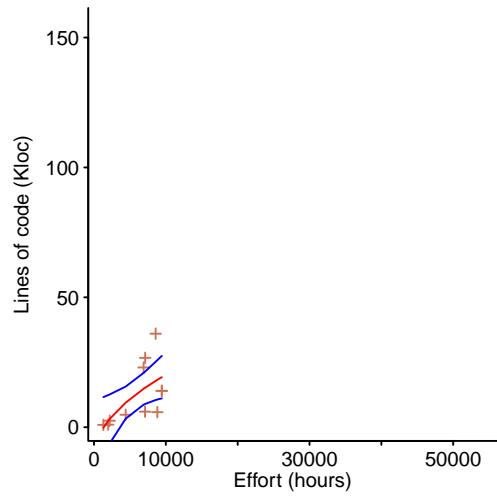


Figure 11.13: Points remaining after removal of overly influential observations, repeatedly applying Cook's distance and Studentized residuals. Data from Fenton et al.<sup>424</sup> [code](#)

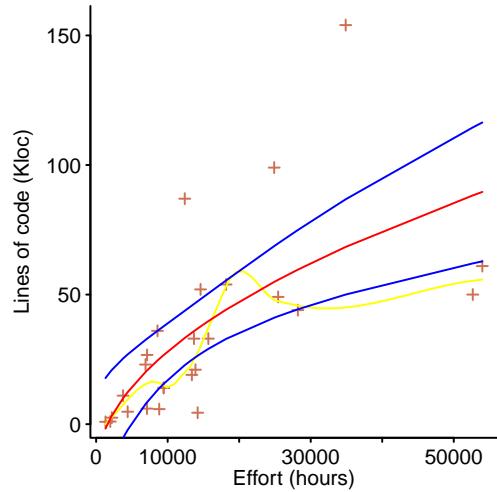


Figure 11.14: Points remaining after removal of overly influential observations, also taking into account the Bonferroni p-value of the Studentized residuals; the line shows the fitted model and 95% confidence interval (loess fit in yellow). Data from Fenton et al.<sup>424</sup> [code](#)

July 31, 2018

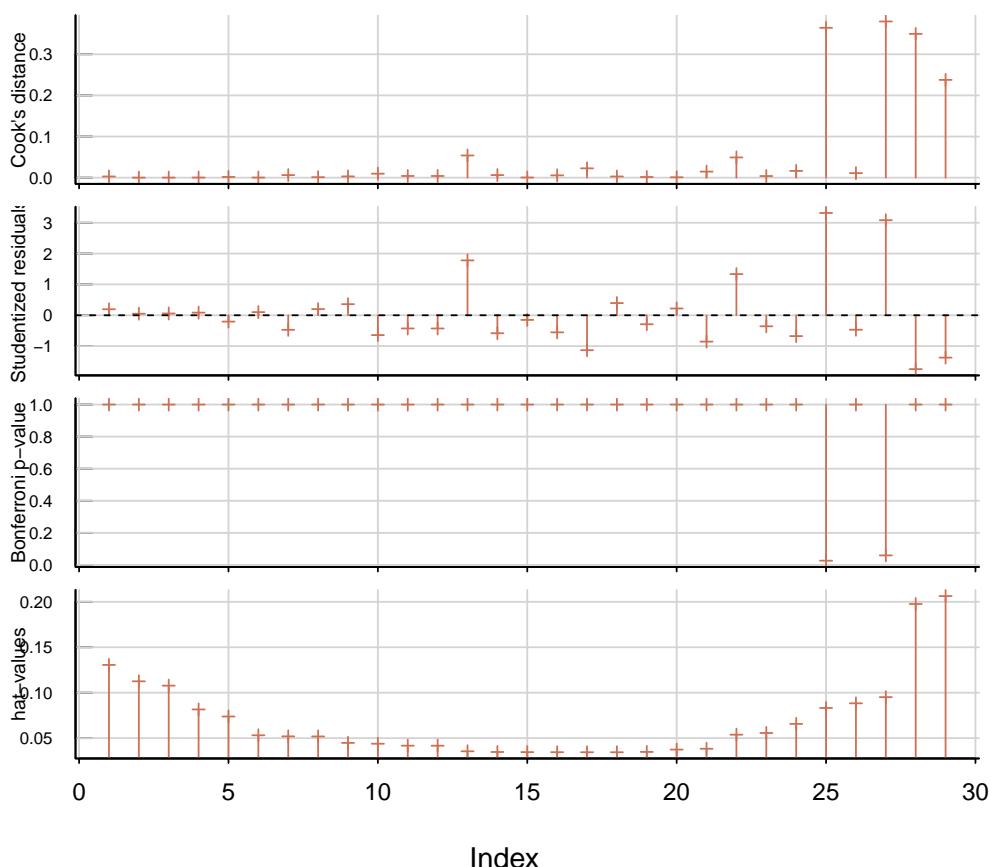


Figure 11.15: `influenceIndexPlot` for the model having the fitted line shown in Figure 11.12. Data from Fenlon et al.<sup>60</sup> [code](#)

```
all_mod=glm(KLoC ~ I(Hours^0.5), data=loc_hour)
influenceIndexPlot(all_mod, main="", col=point_col, cex.axis=0.9, cex.lab=1.0)
```

The top plot shows four data points having a large Cook's distance, but only two of them have a significant corrected p-value (second plot up). These two data points were removed and the process of building a model and calling `influenceIndexPlot` is repeated; this time one point is removed and repeating the process shows that no other data points are worthwhile candidates for removal.

Figure 11.14 shows the results of removing data points having both a high Cook's distance and Studentized residuals whose corrected p-value is below the specified limit.

Outliers are loners, appearing randomly scattered about a plot. When multiple points appear to be following a different pattern than the rest of the data, the reason for this may be a new process driving behavior, or a change of behavior in what went before.

A study by Alemzadeh, Iyer, Kalbarczyk and Raman<sup>19</sup> investigated safety-critical computer failures in medical devices between 2006 and 2011 (as reported by the US Food and Drug Administration). Figure 11.16 shows the number of devices recalled for computer related problems (20-30% of all recalls), binned by two week intervals.

Points that stand out in the upper plot of Figure 11.16 are the two large recall rates in the middle of the measurement interval and recall rates at later dates appearing to increase faster than earlier; plotting a loess fit (green) shows peaks around the two suspicious periods.

The fitted straight line shows a distinct upward trend. Is this fitted line being overly influenced by the middle period or end of period recall rates?

The measurements occur at regular intervals and deleting a measurement in one of these time slots does not make sense; replacing a value with the mean of all observations is one solution for handling this situation.

After replacing two outliers one of the peaks in the fitted loess line is reduced, but there is still a noticeable hump after 2010 (see lower plot). There is little change in the fitted regression model (red and purple lines), showing that the two outliers had little influence. Did a substantive change in the processes driving recalls, or recording of recalls, occur around the start of 2011? Domain knowledge is needed to answer this question.

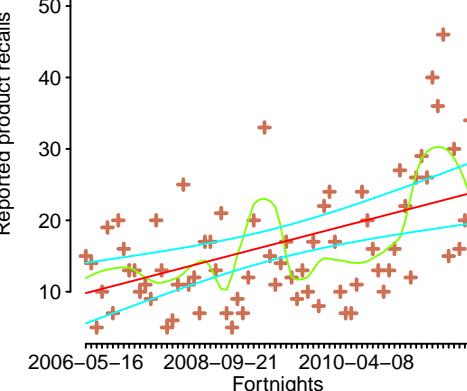


Figure 11.16: Number of medical devices reported recalled by the US Food and Drug Administration, in two week bins. Upper: fitted straight line and confidence bounds, with loess fit (green); Lower: straight line (purple) fitted after two outliers replaced by mean and original fit (red). Data from Alemzadeh et al.<sup>19</sup> [code](#)

In Figure 11.17, the lower plot shows two fitted models one using data up until the end of 2010 and the other using the data after 2010.

This is an example where blindly fitting a straight line to a complete sample produces a misleading model. A change occurred around the end of 2010 that had a significant impact on reported recalls (work is needed to uncover the reason for this change) and fitting data up to the end of 2010 shows a much smaller increase, and perhaps even no increase, in recall rates compared to when measurements after this date are included.

A [change-point analysis](#) of this data is discussed in Section 11.10.

When combining results from multiple studies it is possible for an entire study to be an outlier, relative to other related studies.

A study by Amiri and Padmanabhu<sup>33</sup> analysed the methods used by various other studies to convert between two common methods of counting function points.<sup>xiii</sup> Many of the studies included in the analysis have small sample sizes, include both student and commercial projects, and the function points are sometimes counted by academics rather than industrial developers.

Figure 11.18 shows function points counted using the COSMIC and FPA algorithms (counts made by students have been excluded). Both lines are loess fits, with red used for industry points and blue for academic researchers; the academic line overlays the industry line if one sample (i.e., Cuadtado\_2007) is excluded.

The impact of influential observations on a fitted model can vary enormously, depending on the form on the equation being fitted. Figure 11.19 shows five equations fitted to the Embedded subset of the COCOMO 81<sup>152</sup> data, with the upper plot using the original data and the lower plot the data after three influential observations have been removed from the sample.

In some cases outlier removal has had little impact on the fitted model, while in other cases there has been a dramatic change in the coefficients of the fitted model.

Don't transform variables to reduce the effect of outliers... [developers/74-267-1-PB.R](#)

### 11.2.7 Diagnosing problems in a regression model

The commonly used regression modeling functions will build a model from almost any sample without reporting an error (some functions are so user-friendly they gracefully handle data that produces a singular matrix, an error that is traditionally flagged as it suggests that something somewhere is badly wrong). It is the users' responsibility to diagnose any problems in the model returned.

It is immediately obvious from Figure 11.20 that at least two of the fitted regression lines completely fail to capture the pattern present in the data. This is a famous data set, known as the Anscombe quartet,<sup>46</sup> whose four samples each contain two variables, with each sample having the same mean, standard deviation, Pearson correlation coefficient and are fitted by linear regression with a line having the same slope and intercept.

Problems with a regression model are not always as obvious as the Anscombe quartet case and diagnosing the cause of the problem can be difficult. As always, domain knowledge is very useful for suggesting possible changes to the model.

The difference between the measured value of the response variable and the value predicted by the fitted model is known as the *residual*. Many regression model diagnosis techniques involve the use of the residual. Some of these techniques require a lot more knowledge of the mathematics of regression modeling than is covered in this book.<sup>xiv</sup> Various visualization based techniques are discussed here.

The upper plot in Figure 11.21 shows the residual of the straight line fitted to the Linux kernel growth data analysed [earlier](#). Ideally the residual is randomly scattered around zero and the V-shape seen in this plot is typical of a straight line fitted to values that curve around it (the smallest residual is in the center, where the model fits best, and is greatest at the edges;

<sup>xiii</sup> Function point counting is a technique for estimating development effort by counting the functionality contained in the software requirements specification.

<sup>xiv</sup> It is not obvious that the cost/benefit of learning the necessary mathematics is worthwhile (but it is a good source of homework exercises for students).

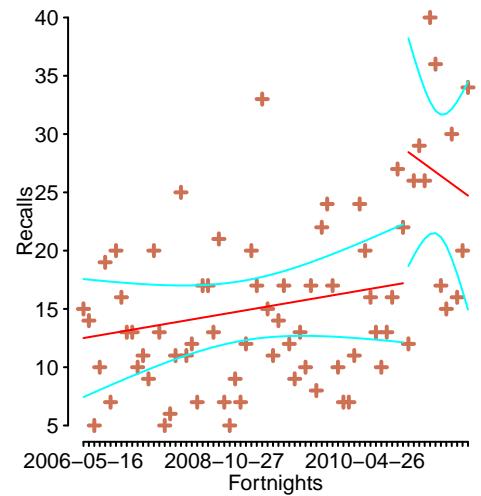


Figure 11.17: Two fitted straight lines and confidence intervals, one up to the end of 2010 and one after 2010. Data from Alemzadeh et al.<sup>19</sup> [code](#)

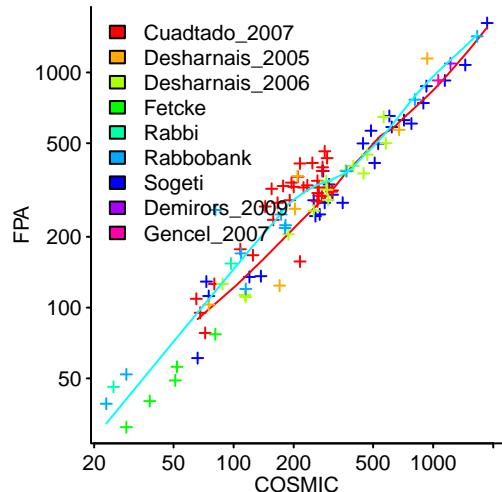
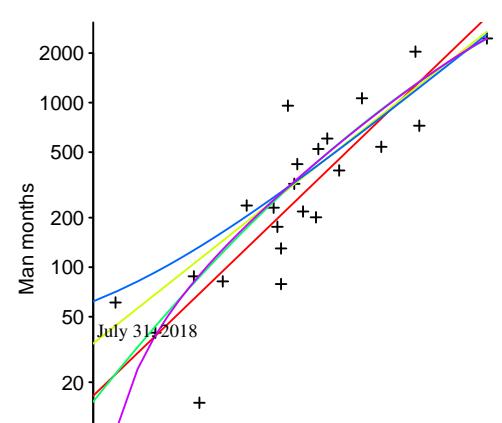
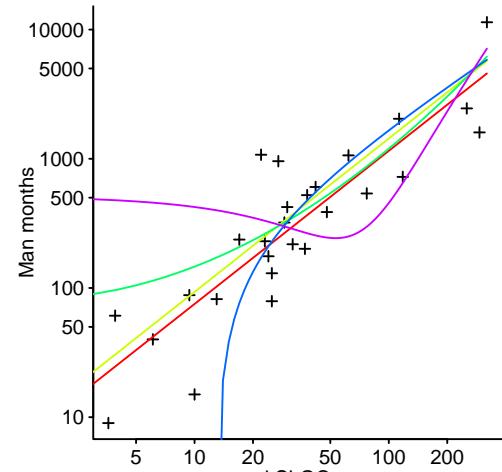


Figure 11.18: Results from various studies of software requirements function points counted using COSMIC and FPA; lines are loess fits to studies based on industry and academic counters. Data from Amiri et al.<sup>33</sup> [code](#)



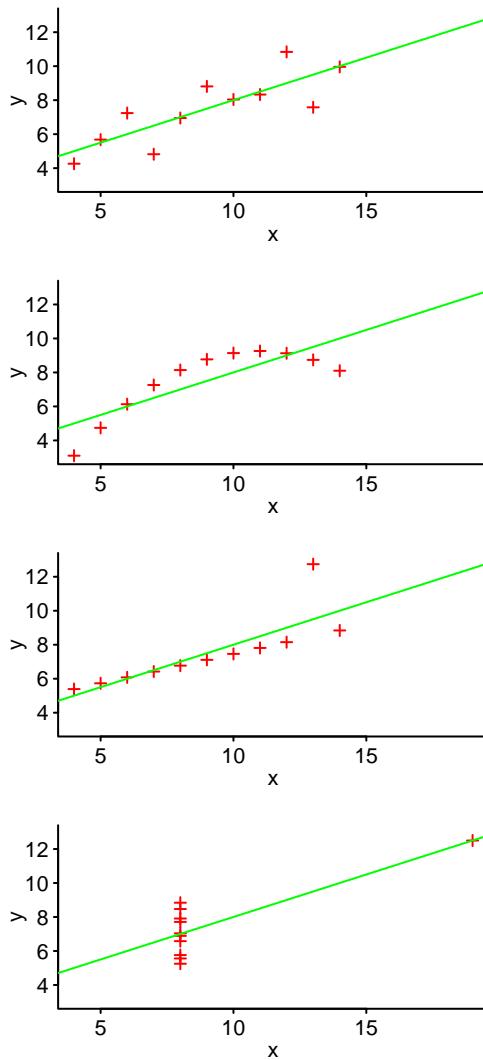


Figure 11.20: Anscombe data sets with Pearson correlation coefficient, mean, standard deviation, and line fitted using linear regression. Data from Anscombe.<sup>46</sup> [code](#)

the smaller peak is a localised change of behavior and may explain why a cubic produces a slightly better fit). This plot shows one of the four diagnostic visualizations produced by `plot` when passed a regression model, as follows:

```
m1=glm(LOC ~ Number_days, data=latest_version)
plot(m1, which=1, caption="", col=point_col)
```

The lower plot of Figure 11.21 shows the original data, straight line fit (red) and loess fit (blue). Both the residual plot and loess fit express the same pattern of curvature around about the straight line fit. Both visualizations have their advantage, the loess line can be drawn before any model is created and the details are easier to extract from a residual plot (e.g., values for the difference).

The mathematics behind linear regression requires that each measurement be independent of all the other measurements in a sample. A common form of dependence between measurements is serial correlation, i.e., correlation between successive measurements. The `durbinWatsonTest` function, in the `car` package, tests a fitted regression model for serial correlation.

A study by Flater and Guthrie<sup>445</sup> measured the time taken to assign a value to an array element in C and C++ using twelve different techniques, some of which checked that the assignment was within the defined bounds of the array (two array sizes were used, large and small); the programs benchmarked were compiled using seven different compiler optimization options.

Figure 11.22 shows the timings from 2,000 executions of one method of assigning to an array element, compiled using `gcc` with the `00` option (upper) and `03` option (lower). The results for `00` show a clustering of execution times for groups of successive measurements.

A Durbin Watson test confirms that the `00` measurements are correlated (see [benchmark/array-durbanwatson.R](#)).

When a non-trivial correlation exists between successive measurements, the appropriate technique to use is time-series analysis, covered in Section 11.10.

## 11.2.8 A model's goodness of fit

This section discusses the various ways in which the error in a regression model is measured and the uses of this value in model selection. The term *goodness of fit* is often used in this context.

When dealing with a single explanatory variable, it is possible to get a good idea of how well a model fits the data through visualization, e.g., by plotting them both. Does the fitted line look correct and how wide are the confidence intervals? However, for data containing more than one explanatory variable, accurate visualizations becomes problematic.

Meeting expectations of behavior is an important model characteristic when building to gain understanding. A model whose behavior goes against expectations is going to have to do a much better job of minimising the differences between the model and the measurement sample than one that is consistent with expectations. In this case error is measured in terms of the difference between expectation and the behavior of the fitted model.

When making predictions, the primary quantity of interest is the accuracy of new predictions, i.e., the amount of expected error in predictions for values that are not in the sample used to build the model. The error structure is also a consideration; is the priority to minimise total error, worse case error, to prefer over estimates to under estimates (or vice versa) or does some complicated weighting (over the range of values that explanatory variables might take) have to be taken into account?

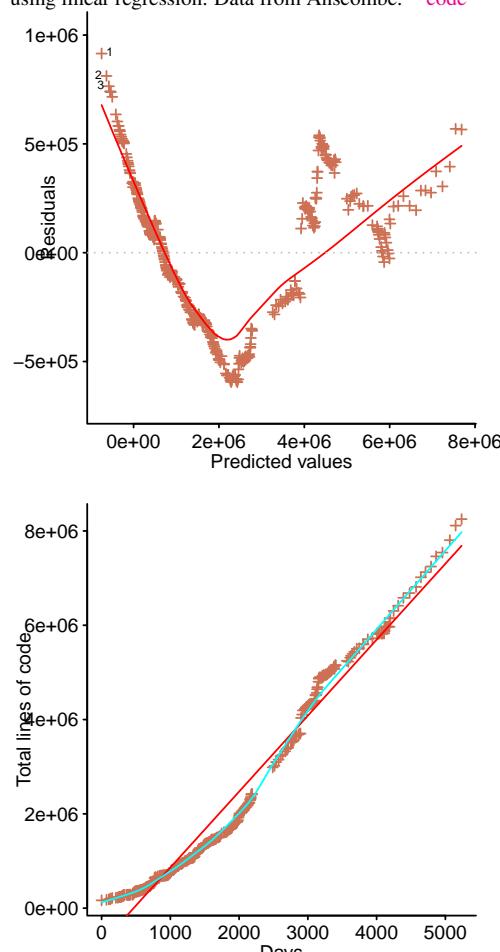
To create a model by fitting it to data, is to create a just so story. The predictions made by a model, outside the range of the data used to build it, are just something to discuss when considering expectations of behavior (which might be derived from a theory of the processes involved in generating the data used to fit the model).

Confidence intervals, see Figure 11.3, provide information about the goodness of fit at every point. The following discussion looks at some of the ways of producing a single numeric value to represent goodness of fit.

The leftover variation in the sample that is not accounted for by the fitted model, the residual, is invariably a component in any equation calculating one or more numbers for the error present in a model. Some of the metrics that readers are likely to encounter include:

v 0.10.0

Figure 11.21: Residual of the straight line fit to the



- deviance is reported by `glm`, in the summary output from a fitted model: null deviance is a measure of the difference between the data and the mean of the data, deviance is a measure of the difference between the data and the fitted model; a relative comparison shows how much better the model performs,
  - R-squared (also known as the *coefficient of determination* and commonly written  $R^2$ ) can be interpreted as the amount of the variance in the data (as measured by the residuals) that is explained by the model. It takes values between zero and one (which has the advantage of being scale invariant) and is a measure of correlation, not accuracy.
- Sometimes the adjusted  $R^2$ ,  $\bar{R}^2$ , is used, which takes into account the number of explanatory variables,  $p$ , and sample size,  $n$ :  $\bar{R}^2 = R^2 - (1 - R^2) \frac{p-1}{n-p}$

- mean squared error (MSE): the mean squared error is the mean value of the square of the residuals and as such has no upper bound (and will be heavily influenced by outliers); root mean squared error (RMSE) is the square-root of MSE.

The following equation shows how MSE and  $R^2$  are related:  $R^2 = 1 - \frac{MSE}{\sigma^2}$

- mean absolute error (MAE): the mean absolute error is the mean value of the absolute value of the residuals. This measure is more robust in the presence of outliers than MSE.

Apart from R-squared metric, the metrics listed (plus AIC) are scale dependent, e.g., mapping measurements from centimeters to inches changes their value; transforming the scale (e.g., taking logs) will also change values.

The choice of a metric is driven by what information is available and what model characteristics are considered important (e.g., how important is outlier handling). In a competitive situation people might not be willing to reveal details about their model and so any metric has to be made on predictive accuracy (e.g., models builds provide the predictions made by their model to a test data set).

R-squared is the only scale invariant metric and provides an indication of how much improvement might be possible over an existing model.

It is possible for the coefficients of a fitted model to be known with a high degree of accuracy and yet for this model to explain very little of the variance present in the data, and for there to appear to be little chance of improving on the model given the available data.

The Ultimate Debian Database project<sup>[1376](#)</sup> collects information about packages included in the Debian Linux distribution. Figure 11.23 shows the age of a packaged application plotted against the number of systems on which that application is installed, for 14,565 applications in the "wheezy" version of Debian. Also, see Figure 6.2.

The fitted linear model (red line hidden by the 95% confidence interval in green overwriting it; loess fit in blue) has a very low p-value, a consequence of the large number of, and uniform distribution of, data points. The predictive accuracy of this model is almost non-existent, the only information it contains is that older packages are a little more likely to be installed than younger ones.

A study by Jørgensen and Sjøberg<sup>[721](#)</sup> investigated developers ability to predict whether any major unexpected problems would occur during a software maintenance task. Building a regression model using the available measured attributes showed that lines of code was the only explanatory variable having a p-value less than 0.05. However, only 3.3% of the variance in the response variable was explained by lines of code; so while the explanatory variable was statistically significant, its practical significance was negligible (see [maintenance/10.1.1.37.38.R](#)).

### 11.2.9 Low signal-to-noise ratio

Sample measurements sometimes contain a large amount of noise relative to the signal that is present, i.e., a low signal-to-noise ratio. Fitting a model to data having a low signal-to-noise ratio can be difficult because many equations do an equally good job.

The top row of Figure 11.24 shows data generated from a quadratic equation containing noise and two fitted models. The following equation was used to generate the two sets of data:

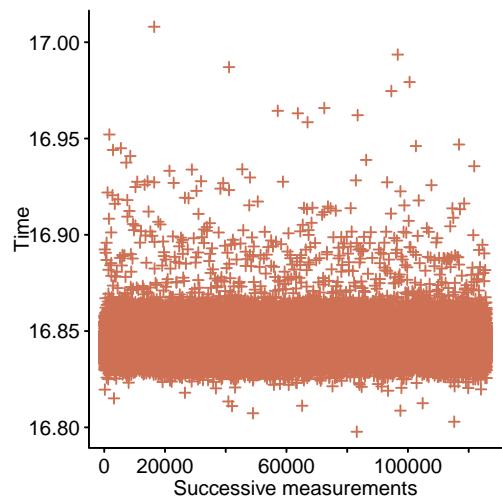
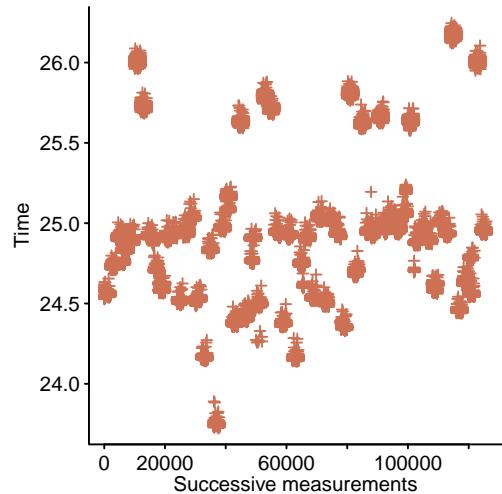


Figure 11.22: Array element assignment benchmark compiled with gcc using the 00 (upper) and 03 (lower) options (measurements were grouped into runs of 2,000 executions). Data from Flater et al.<sup>[445](#)</sup> [code](#)

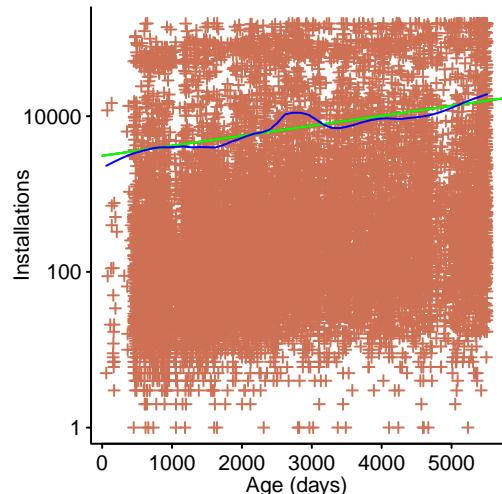


Figure 11.23: Number of installations of Debian packages against the age of the package, plus fitted model and loess fit. Data from the "wheezy" version of the Ultimate Debian Database project.<sup>[1376](#)</sup> [code](#)

$$y = x^2 + K \times (5 + rnorm(length(x)))$$

where:  $K$  was  $10^3$  (left column) or  $10^2$  (right column).

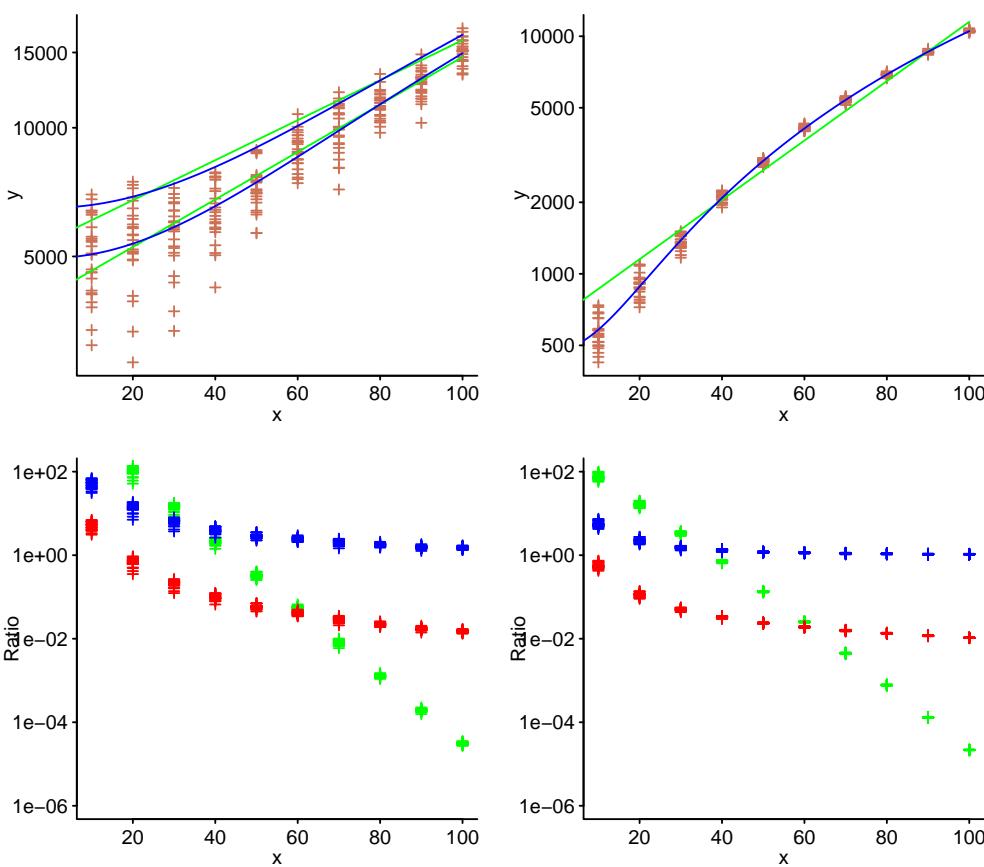


Figure 11.24: Quadratic relationship with various amounts of added noise fitted using a quadratic and exponential model. [code](#)

It is not possible to tell by looking at the top left plot whether a quadratic (blue) or an exponential (green) is a better fit; the output from `summary` is not much help (see [regression/noisy-data.R](#)). The top right plot contains less noise and it is easier to see that the exponential fit does not follow the data as well as the quadratic.

Sometimes the peaks (or troughs) can be a better indicator of the shape of the data. The top left plot shows a quadratic and exponential fit to the three largest values at each  $x$ -value. The fit does not seem to reduce the uncertainty in this case.

The *ratio test* is a technique that can help rule out some possible model equations. If  $f(x)$  is the function being fitted to the data and this data was generated by the function  $g(x)$ , the ratio  $\frac{g(x)}{f(x)}$  will converge to a constant as  $x$  becomes small/large enough such that the signal dominates the noise.

The bottom row in Figure 11.24 show ratio tests for quadratic (blue), cubic (red) and exponential (green) equations. The exponential equation shows no sign of converging to a constant, while a quadratic is closer to doing this than a cubic (which can be ruled out because it does a poor job of fitting the data).

The ratio test rules out an exponential equation being a good model for the sample data.

A study by Vasilescu, Serebrenik, Goeminne and Mens<sup>[1395](#)</sup> investigated contributions to the Gnome ecosystem, from the point of view of workload (measured by counting the number of file touches, e.g., commits), breaking it down by projects, authors and number of activity types (e.g., coding, testing, documentation, etc).

The upper plot of Figure 11.25 shows, for individual authors, workload and the number of activity types they engaged in. There is a large amount of noise in the data. The lower plot of Figure 11.25 shows a ratio test, with an exponential equation failing to level off, the linear equation slowly growing and the quadratic looking like it is trying to grow.

Perhaps the situation becomes clearer with more activity types, but none of the commonly occurring equations looks like they may be good fits.

Figure 11.25: Author workload against number of activity

types per author (upper) and ratio test (lower). Data from

### 11.2.10 Weighting data

Ho hum, some data where weighting has a meaningful interpretation...

### 11.2.11 Sharp changes in a sequence of values

When the processes generating the measured values changes, the statistical properties of the post-change sequence of values may change. The point at which the statistical properties of a value sequence changes is known as a *change-point*.

The changepoint package includes functions for performing change-point detection of the mean, variance and both mean and variance of a sequence of values. The cpt.mean function checks for significant shifts in the mean value; the two options are method="AMOC" (At Most One Change; other values support searching for a specified maximum number of changes, with method="AMOC" selecting what it considers to be the optimum number of changes) and test.stat="Normal" (assume the measurement error has a Normal distribution).

An earlier analysis of electronic device recall frequency suggested that a significant shift in processes driving recalls occurred at the end of 2010. Figure 11.26 shows the output from the following calls to cpt.mean:

```
change_at=cpt.mean(as.vector(t2))
plot(change_at, col=point_col,
     xlab="", ylab="Reported product recalls\n")

change_at=cpt.mean(as.vector(t2), method="PELT")
plot(change_at, col=point_col,
     xlab="Fortnights", ylab="Reported product recalls\n")
```

See [regression/hpc-read-write.R](#) for an example of detecting changes in variance and changes in both mean and variance.

The existence of a change-point is not always obvious and when dealing with a few discrete measurement points change-point analysis might not be able to provide a reliable answer. As always, a loess curve can provide useful information.

A study by Berger, She, Lotufo, Wąsowski and Czarnecki<sup>129</sup> analysed the variability models for 13 open source projects. Figure 11.27 shows, for systems containing a total number of possible features, the number of these features depending on 1, 2, 3, etc. build-flags (some optional features require other optional features to be enabled before they can be enabled).

The loess curve (in red) is flat and then suddenly jumps and appears to flatten off again. A fitted regression line (in green) appears to have been shifted up by the jump in values. Perhaps the processes driving optional features in the few larger systems that were measured are different from the smaller systems.

A sequence of values containing discontinuities can be fitted by multiple regression models, one each over each interval between discontinuities. However, it is possible to build a single model that contains the discontinuity information.

Figure 11.28 shows a distinct change in the pattern of the sales volume of 4-bit microprocessors (green). Straight lines have been fitted to the two periods before/after April 1998 (red), with the yearly sales cycle modeled with a single sine wave (blue).

The technique for building a model to handle discontinuous patterns of behavior makes use of an interaction between the explanatory variable, date, and a dummy variable whose 0/1 value depends on date relative to the discontinuity point. The code for the straight line model (in red) is:

```
# discontinuity point
y_1998=as.Date("01-04-1998", format="%d-%m-%Y")

p4=glm(bit.4 ~ date*(date < y_1998)+date*(date >= y_1998), data=proc_sales)
```

and the summary output is: [code](#)

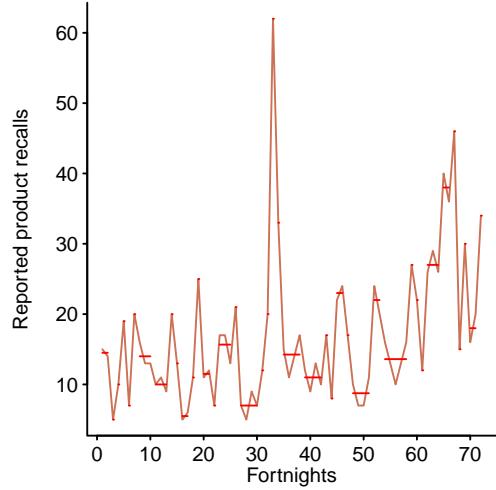
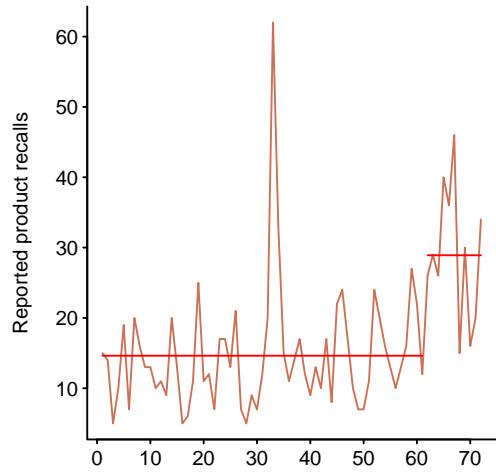


Figure 11.26: Change-points detected by cpt.mean, upper using method="AMOC" and lower using method="PELT". Data from Alemzadeh et al.<sup>129</sup> [code](#)

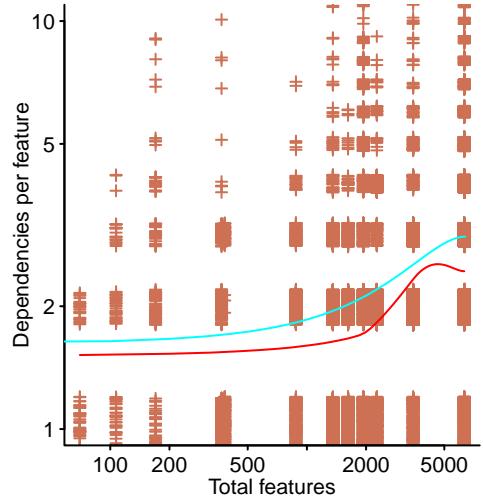


Figure 11.27: Number of flags (y-axis jittered) used to control the selection of optional features in system containing a total number of features, loess curve (red), regression line (green). Data from Berger et al.<sup>129</sup> [code](#)

```

Call:
glm(formula = bit.4 ~ date * (date < y_1998) + date * (date >=
y_1998), data = proc_sales)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-19756.9 -6372.8 -558.7  6533.2 19086.4 

Coefficients: (2 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       7.050e+04  5.802e+04   1.215   0.2265  
date            7.072e-01  5.368e+00   0.132   0.8954  
date < y_1998TRUE -8.357e+04  5.873e+04  -1.423   0.1572  
date >= y_1998TRUE        NA        NA     NA     NA      
date:date < y_1998TRUE  1.045e+01  5.466e+00   1.912   0.0581 .  
date:date >= y_1998TRUE        NA        NA     NA     NA      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 80003408)

Null deviance: 2.0763e+10 on 131 degrees of freedom
Residual deviance: 1.0240e+10 on 128 degrees of freedom
AIC: 2782.6

Number of Fisher Scoring iterations: 2

Sales follow a seasonal trend that can be approximated using a single 12-month frequency
sine wave and adding this to the straight line model:

season_p4=glm(bit.4 ~ date*(date < y_1998)+date*(date >= y_1998)+
              I(sin(rad_days))+I(cos(rad_days)),
              data=proc_sales)

```

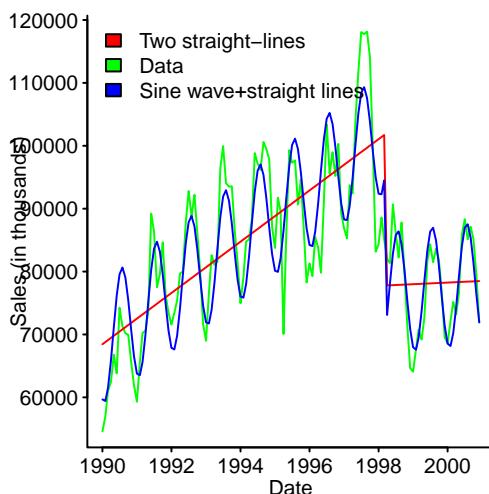


Figure 11.28: Monthly unit sales (in thousands) of 4-bit microprocessors. Data kindly supplied by Turley.<sup>1369</sup>  
code

### 11.3 Moving beyond the default Normal error

Some measurements in software engineering have properties that do not meet the requirements necessary for the mathematics behind glms default argument values to work. These measurements often have one or more characteristics that require non-default argument values to be passed to `glm`, including the response variable having:

- values that span several orders of magnitude,
- values that can never go below zero, e.g., count data,
- values that can never go above some maximum value, e.g., a percentage can never be greater than one hundred.

By default, `glm` assumes that measurement error has a Normal distribution (also known as a \_Gaussian distribution<sup>xv</sup>). Figure 11.29 shows a fitted regression line with four of the data points (in red); the colored Normal curves over each point represents the probability distribution of the measurement error that is assumed to have occurred for that measurement (the center of each error curve is directly above the fitted line at each sample value of the explanatory variable).

In calls to `glm`, the `family` argument has the default value `family=gaussian(link="identity")`, which can be shortened in code to `family=gaussian` because `link="identity"` is the default link function for `gaussian`.

The Normal distribution includes negative values and when a measurement cannot have a negative value, using an error distribution that allows negative values to occur can distort the fitted model. One possible alternative is the Poisson distribution, which is zero for all negative values. The following call to `glm` specifies that the measurement error has a Poisson distribution:

<sup>xv</sup> This book uses the term Normal because it appears to be more widely used.

```
a_model=glm(a_count ~ x_measure, data=some_data, family=poisson)
```

After Normal, the Poisson and Beta distributions are the most common error distributions appearing in this book.

Calling `glm` using a non-default value for the `family` argument requires knowing more about the mathematics behind generalised regression model building. The equation actually being fitted by `glm` is:

$$l(y + \varepsilon) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

which differs from the one given at the start of the chapter in having  $l(y + \varepsilon)$  on the left-hand-side, rather than  $y$ . This  $l$  is known as the *link function* and for the Normal distribution is the identity function (which leaves its argument unmodified, which means the equation ends up looking like the equation given at the start of the chapter).

Once a regression model is fitted, the value of the response variable is calculated from:

$$y = l^{-1}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n) - \varepsilon$$

where:  $l^{-1}$  is the inverse of the link function used, e.g., the inverse of log is  $e$  raised to the appropriate power.

Every error distribution has what is known as a *canonical link* function, which is the function that pops out of the mathematical analysis for that distribution. By default, `glm` uses the canonical link function for each error distribution, and allows some alternatives to be specified. The canonical link function for the Poisson distribution is `log`.

When the link function is not `identity`, prediction values and confidence intervals need to be mapped as follows

```
a_pred=predict(a_model, se.fit=TRUE)
inv_link=family(a_model)$linkinv      # get the inverse link function

lines(x_values, inv_link(a_pred$fit)) # fitted line
# confidence interval above and below
lines(x_values, inv_link(a_pred$fit+1.96*a_pred$se.fit))
lines(x_values, inv_link(a_pred$fit-1.96*a_pred$se.fit))
```

The following sections analyse measurements having characteristics that may require the use of various error distributions and link functions.

### 11.3.1 Count data

Count data has two defining characteristics, it is discrete and has a lower bound of zero. The discrete distribution only taking on non-negative values, supported by `glm`, is the Poisson distribution.

In practice, when measurement values are sufficiently far away from zero (where far may be more than 10) there is little difference between models fitted using the Normal and Poisson distributions. For measurements closer to zero the main difference between models fitted using different distributions is in the confidence intervals (which are usually not symmetric and may be larger/smaller).

The canonical link function for the Poisson distribution is `log` and the following two calls are equivalent:

```
p_mod=glm(y ~ x, data=sample, family=poisson)
p_mod=glm(y ~ x, data=sample, family=poisson(link="log"))
```

The `log` link function means that the equation being fitted is actually:

$$y = e^{\alpha + \beta x}$$

To fit the equation:  $y = \alpha + \beta x$ , using a Poisson error distribution, the `identity` link function has to be used, as follows:

```
p_mod=glm(y ~ x, data=sample, family=poisson(link="identity"))
```

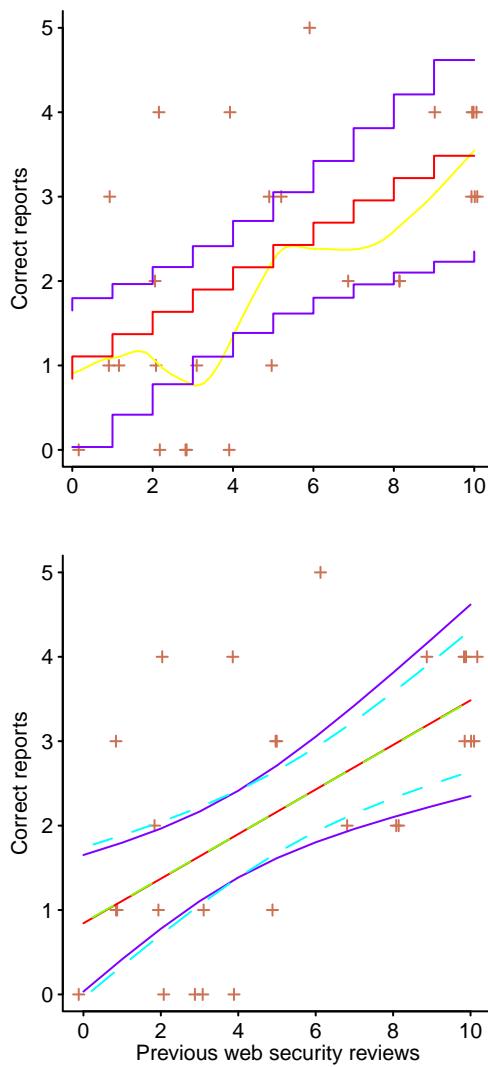


Figure 11.30: Number of vulnerabilities detected by professional developers with web security review experience; upper: technically correct plot of model fitted using a Poisson distribution, lower: easier to interpret curve representation of fitted regression models assume error has a Poisson distribution (continuous lines) or a Normal distribution (dashed lines). Data extracted from Edmundson.<sup>375</sup> code

A study of the effectiveness of security code reviews by Edmundson, Holtkamp, Rivera, Finifter, Mettler and Wagner<sup>375</sup> asked subjects, professional developers with web security review experience, to locate vulnerabilities in web code. The number of vulnerabilities found can only ever be a non-negative integer value and in this study were single digit values.

The values fitted to a discrete distribution consists of a series of discrete steps, as the upper plot of Figure 11.30 shows (fitted line and 95% confidence intervals). While this plot is technically correct, it is ambiguous: are the values specified by the top left edge or the bottom right edge of the staircase?<sup>xvi</sup> Plots using continuous lines are easier for readers to interpret and these are used in this book.

The continuous lines in the lower plot of Figure 11.30 were fitted using the argument `family=poisson(link="identity")`, while the dashed lines were fitted using `glm` default argument values.<sup>xvii</sup>

The two fitted lines are virtually identical (the green dashed line is drawn over the continuous red line), the 95% confidence intervals are different. This pattern of behavior is very common unless the response variable has many values near zero.

When fitting models containing multiple explanatory variables (discussed later) and a response variable containing count data, it can be more difficult to see the difference between using a Poisson and Normal distribution for the errors. Using a Poisson distribution may be an inconvenience, but it removes uncertainty and is always worth trying.

Sometime a Poisson error distribution is used because a `log` link function is required to model an additive error.

The Negative Binomial distribution is probably the second most commonly occurring count distribution. A study by Jones<sup>699</sup> included counting the number of break statements in C functions. A break statement can occur zero or more times within a loop or switch statement, and these statements can occur zero or more times within a function definition. One process for generating a Negative Binomial distribution is to select values from multiple Poisson distributions, whose mean has a Gamma distribution. Figure 11.31 shows the number of functions containing a given number of break statements, along with a fitted Negative Binomial distribution.

The `gamlss` package supports a wide variety of probability distributions, including the NBI distribution used in the following call to the `gamlss` function:

```
breaks=rep(j_brk$occur, j_brk$breaks)
nbi_bmod=gamlss(breaks ~ 1, family=NBI)

plot(function(y) max(jumps$breaks, na.rm=TRUE)*      # Scale probability distribution
      dNBI(y, mu=exp(coef(nbi_bmod, what="mu")),
            sigma=exp(coef(nbi_bmod, what="sigma"))),
      from=0, to=30, log="y", col=pal_col[1],
      xlab="breaks", ylab="Function definitions\n")
points(jumps$occur, jumps$breaks, col=pal_col[2])
```

While zero is a common lower bound, other values are sometimes encountered. When the lower bound is one, a zero truncated error distribution can be used. The `vglm` function in the `VGAM` package supports a wide variety of error distributions functions, including zero-truncated...

Calls to `vglm` follows the same pattern as calls to `glm`, except a wider variety of distribution families are supported (`pospoisson` is the zero-truncated Poisson distribution).

```
feat_mod=vglm(all ~ total, data=q, family=pospoisson())
```

The `quasipoisson` function...

Modeling zero inflated data... the `mhurdle` package... the `gamlss` package... Zero inflated poisson ...`zipoisson()`

<sup>xvi</sup> The choice is selectable via the `type` argument to `plot/lines`.

<sup>xvii</sup> To achieve an acceptable p-value, three outliers were removed.

### 11.3.2 Continuous response variable having a lower bound

Some continuous response variable measurements have a lower bound of zero, e.g., length or time measurements. The continuous distribution only taking on non-negative values, supported by `glm`, is the Gamma distribution.

In practice, when most measurement values are sufficiently far away from zero (where far away could be a large single digit value) there is very little difference between models fitted using the Normal and Gamma distributions. For measurements closer to zero the main difference between models fitted using different distributions is in the confidence intervals (which are usually not symmetric and may be larger/smaller).

The canonical link function for the Gamma distribution is `inverse` and the following two calls are equivalent:

```
G_mod=glm(y ~ x, data=sample, family=Gamma) # Yes, capital G
G_mod=glm(y ~ x, data=sample, family=Gamma(link="inverse"))
```

The `inverse` link function means that the equation being fitted is actually (the `identity` link function is supported for this distribution):

$$y = \frac{1}{\alpha + \beta x}$$

Figure 11.32 comes from a code review study (discussed in Section 13.2) and shows meeting duration when reviewing various amounts of code. Meeting duration must be greater than zero and a Gamma error distribution is assumed to apply (the data has a linear relationship and the identity link function is used). The green line is the model fitted using a Gaussian error distribution.

The data contains a few points with high leverage and the loess fit suggests that there may be a change-point, so a more involved analysis is probably necessary.

### 11.3.3 Transforming the response variable

When plotting sample points, values along one or both axes are sometimes transformed to compress or spread out the points, for the purpose of improving data visualization, e.g., using a log scale.

A regression model is fitted to a pattern (i.e., an equation) and if a visible pattern of behavior is present in a plot using transformed axis, it is worth investigating a model that uses similarly transformed values.

Applying a non-linear transform to the response variable changes its error distribution and a regression model built using this transformed response variable may not be a natural fit to the processes that generated the measurements. Explanatory variables are assumed not to contain any error and transforming them does not change this assumption.

For example, in the following regression model the error,  $\epsilon$ , is additive:

$$y = \alpha + \beta x + \epsilon$$

while fitting a log-transformed response variable:

$$\log y = \alpha + \beta x + \epsilon$$

produces a model where the error is multiplicative, i.e., the error present is a percentage of the measured value:

$$y = e^{\alpha + \beta x} e^\epsilon$$

The error in a model fitted using a log link function is additive, because the equation fitted is:

$$\log(y + \epsilon) = \alpha + \beta x$$

which becomes:

$$y = e^{\alpha + \beta x} + \epsilon$$

If the response variable is transformed, the decision on whether to transform it directly or via a link function is driven by whether the error is thought to be additive or multiplicative. As always, knowledge of the processes that produced the measured values is drive the decision.

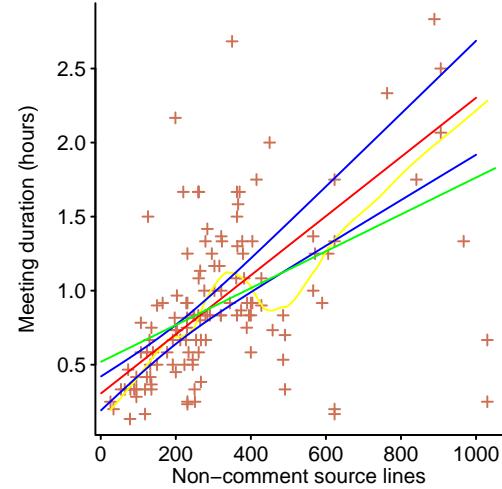


Figure 11.32: Code review meeting duration for a given number of non-comment lines of code; fitted regression model, assuming errors have a Gamma distribution (red, with confidence interval in blue) or a Normal distribution (green). Data from Porter et al.<sup>1104</sup> [code](#)

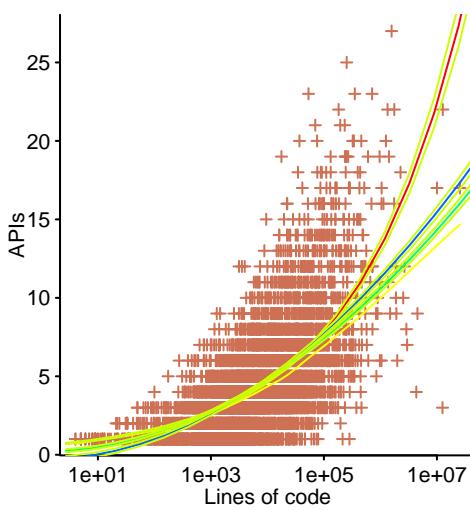


Figure 11.33: Number of APIs used in Java programs containing a given number of lines and three fitted models. Data from Starek.<sup>1301</sup> [code](#)

A study by Starek<sup>1301</sup> investigated API usage in Java programs. Figure 11.33 shows the number of APIs used in Java programs containing a given number of lines of code. Many programs use a few APIs, and the number of APIs is not large, suggesting that a Poisson error distribution may be applicable; the range in the number of APIs used does not suggest a log scale.

The following three models were fitted (plot line color shown with 95% confidence bounds sharing the same color and a yellow loess fit peaking through underneath):

```
ap_mod=glm(APIs ~ log(Size), data=API, family=poisson) # Red
ap_id_mod=glm(APIs ~ I(log(Size)^2), data=API,
               family=poisson(link="identity"),
               start=c(1, 1)) # Green
ag_mod=glm(APIs ~ I(log(Size)^2), data=API) # Blue
```

Which of these models is the best description of the processes that produced the measurements? Perhaps none of them.

Choosing a model based on one of the available goodness of fit metrics is one possibility. Ideally, the choice should be based on a theory, or hypothesis, that explains the distribution of usage, along with measurements of other program characteristics capably of being combined into a model that explains more of the variance in the data.

Except for when many sample values are close to zero, the Poisson distribution is a good enough approximation to the Gaussian distribution that it can be substituted when a log link function is required and the response variable take's integer values (see [sourcecode/ICSE2010-cpp/ICSE2010-cpp.R](#)). When the response variable contains many non-integer values (i.e., too many to ignore), then a log link function can be specified: `family=gaussian(link="log")` (see [regression/putnam-MTTF.R](#)).

One advantage of log transforming a response variable is that it reduces the influence of outlier values (because the range of values is compressed). Figure 11.34 illustrates the impact of removing one highly influential value from the data used to fit a model using a log link function (green lines) and a model fitted to a log transformed response variable (red lines).

The visual appearance of outliers and influential observations plotted using log axis can be deceiving, i.e., they may not appear to be that far removed from the general trend. As always, assumptions based on visual appearance need to be checked using numerical methods.

The `robustbase` package includes the `glmrob` function for robust GLM model fitting, which is not always as robust as desired and manual help may be required (see [regression/a174454-reg.R](#)).

If the only available practical to-use regression modeling technique assumes the error in the response variable has a Normal distribution, then there is an incentive to transform the response variable so that it has this characteristic. When a computer, and the necessary software, is available to do the calculation, there is no need to transform the response variable to give it a normally distributed error.

The `boxcox` function in the `MASS` package and the `powerTransform` function in the `car` package suggest the transform that is most likely to produce a response variable whose error has a Normal distribution.

A traditional approach to simplifying a problem is to map a continuous variable to a number of discrete values (e.g., small/medium/large). Throwing away information may simplify a problem, but the cost is a considerable loss of statistical power and residual confounding.<sup>1183</sup> Using a computer removes the need to simplify in order to reduce the manual effort needed to perform the analyse.

See [regression/melton-statics.R](#) for an example where building a regression model provides a lot more information about the characteristics of the continuous data compared to mapping values to large/small and running a chi-squared test.

Adjusting a model to handle uncertainty in explanatory variables, when the model contains a multiplicative error, requires specifying the measurement error for every value of an explanatory variable. The following option assigns a 10% error `measurement.error=maint$lins_up/10`.

A study by Jørgensen<sup>709</sup> investigated maintenance tasks and obtained developer effort and code change data. Figure 11.35 shows the effort (in days) and number of lines inserted and updated for 89 maintenance tasks. The original fitted regression line is in red and the SIMEX adjusted line is in blue. The following is the call to `simex`:

```
maint_mod=glm(EFFORT ~ lins_up, data=maint,
               family=gaussian(link="log"), x=TRUE, y=TRUE)

y_err=simex(maint_mod, SIMEXvariable="lins_up",
            measurement.error=maint$lins_up/10, asymptotic=FALSE)
```

### 11.3.4 Binary response variable

When the response variable can only take one of two possible values, e.g., (false, true) or (0, 1), it has a binomial distribution. If the response variable, as the explanatory variable increases (or decreases), switches from 0 to 1 (or 1 to 0) and then always has that value for further increases in the explanatory variable, there is no need to build a regression model (simply find the switch point). When the response variable can have two possible values over some range of the explanatory variable, regression modeling will fit an equation that minimises the residual error.

The data in Figure 11.36 is continuous on the x-axis and has two discrete values on the y-axis. While it is possible to fit a regression model using a straight line, it is difficult to interpret this line as a model of the processes that generated the data. A Logistic equation starts low (or high) and eventually the high (or low) values dominate and it switches.

[developers/74-267-1-PB.R](#) from Höfer<sup>625</sup>... or [economics/upgrade-languages.R](#)

The canonical link function for the Binomial distribution is `logit` and the following two calls are equivalent:

```
b_mod=glm(y ~ x, data=sample, family=binomial)
b_mod=glm(y ~ x, data=sample, family=binomial(link="logit"))
```

The equation for the `logit` link function is:

$$\log \frac{y}{1-y} = \alpha + \beta x$$

where the response has the form of log-odds ratio.

The predicted values returned by `predict`, from a fitted binomial model, are in the range 0...1. The user has to make a decision about where to divide this continuous range, with predicted values on one side of the division treated as zero and all other values as one. A simple approach is to treat predictions greater than 0.5 as one and everything else as zero, while a more sophisticated approach looks at the distribution of predictions and makes an informed trade-off between true positive (in this context also known as *recall* and *hit rate*) and accuracy (i.e., false positive rate).

A ROC curve (receiver operating characteristics; named after a technique used to measure the performance of radio receivers) is a visualization technique for showing the trade-offs between two rates, e.g., true positive rate and false positive rate; it is a popular technique for displaying the trade-offs from predictions returned by machine learning models.

The `ROCR` package supports the creation and plotting of ROC curves.

The pair of columns in Table 11.1 is an example, from training data, of the impact of selecting particular cut-off values for distinguishing between true/false (for 10 data points). At a cut-off of 0.9 one correct prediction, a true positive, is made (20% of the available true responses), at a 0.81 cut-off another correct prediction occurs, while at 0.72 an incorrect prediction, a false positive, is made (at this cut-off the response rate for correct predictions is 40% and 20% for incorrect predictions).

t	t	f	t	f	t	f	t	f	f
0.90	0.81	0.72	0.60	0.53	0.44	0.39	0.28	0.16	0.09

Table 11.1: Example list of actual prediction outcome occurring at various cut-off values. [code](#)

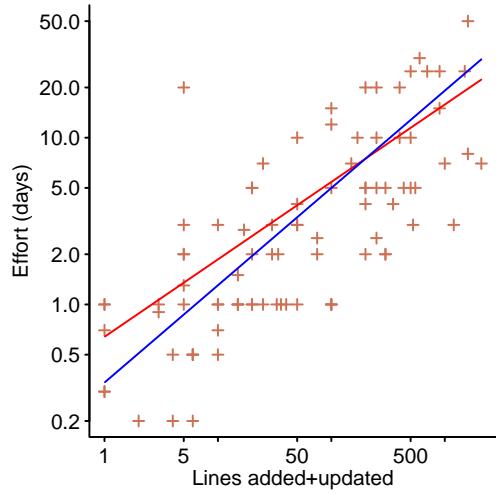


Figure 11.35: Maintenance task effort and lines of code added+updated, with fitted regression model (red) and SIMEX adjusted for 10% error (blue). Data from Jørgensen.<sup>709</sup> [code](#)

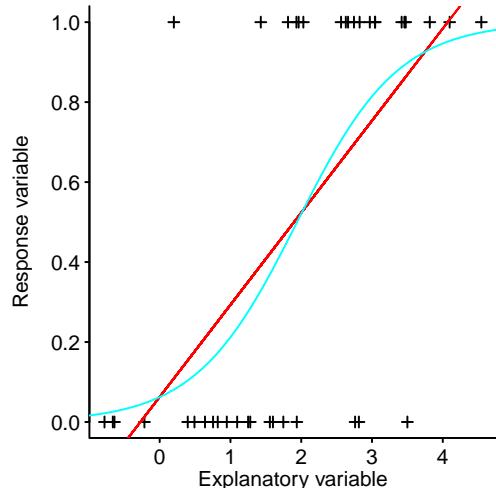


Figure 11.36: Regression modeling 0/1 data with a straight line and a logistic equation. [code](#)

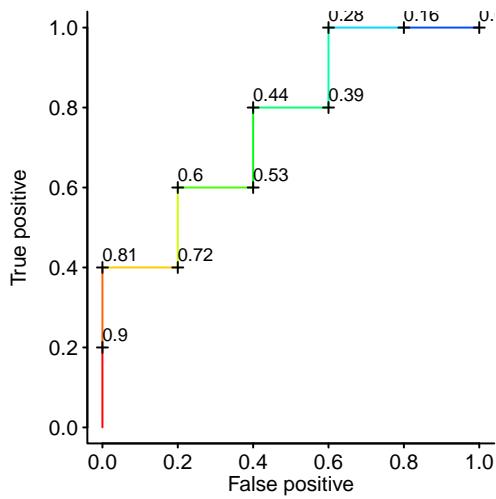


Figure 11.37: ROC curve for the data listed in Table 11.1.  
code

Figure 11.37 shows the ROC curve for this data...

AUC (Area Under the Curve) and select a cut-off value that provides the best trade-off between false-positive and....

Comparing the fitted model coefficients by plotting them using:

```
library("sjPlot")
...
sjp.glm(proc_mod)
```

In Figure 6.32... [reliability/fuzzer/fuzzer-mod.R](#) is a relatively complicated model... need something simpler...

### 11.3.5 Multinomial data

When discrete response variables take on more than two values, they have a *multinomial* distribution.

**nominal** : When a response variable can take  $N$  distinct values and  $\pi_i$  is the probability of the  $i^{th}$  value occurring then the baseline-category logit model, with one explanatory variable  $x$ , is (with the sum over all  $\pi$  equals one):

$$\log \frac{\pi_n}{\pi_N} = \alpha_n + \beta_n x$$

for  $n = 1, \dots, N - 1$ .

Building a model results in  $N - 1$  equations with separate coefficients for each.

The `mlogit` package supports the building of multinomial logit...

**ordinal** : fitting a logit model to each pair of adjacent values does not make use of all the information present and the logit equation can be extended to make use of the ordering information...

The model building process pairs each value of the multinomial response variable with a baseline value from that variable (where the baseline value is the last value *baseline-category logit*...)

The cumulative probability for  $Y$  is the probability that  $Y$  is at or falls below a given value:

$$P(Y \leq i) = \pi_1 + \pi_2 + \dots + \pi_N$$

The *cumulative logits* treats the  $P(Y \leq i)$  as the response variable in a logit model building process and creates a separate model for each value of  $i$ .

The `ordinal` package supports the building of *cumulative link models*, also known as *ordinal regression models*.

[developers/LuthigerJungwirth.R...](#)

Perhaps use CART when there are several categorical variables, early splits involving these variables may suggest that multiple models be created...

### 11.3.6 Rates and proportions response variables

When dealing with a response variable representing a rate or proportion there is a fixed lower and upper bound, often zero and one or 100. Measurements within an interval often share particular characteristics: they exhibit more variation around the mean and less variation towards the lower and upper bounds,<sup>xviii</sup> and they may have an asymmetrical distribution. These characteristics can be accommodated by the *Beta distribution*; a regression model where the response variable has a Beta distribution is known as a *Beta regression model*.

The `betareg` package contains functions that support building a Beta regression model. When fitting basic models, calls to `betareg` have the same form as calls to `glm`; both functions include more sophisticated options that are not supported by the other.

Figure 11.38 shows fitted curves from a beta regression model (red) and a call to `glm` (blue); the study from which the data is taken is discussed elsewhere, see Figure 6.47. The equation

<sup>xviii</sup> The measurement sample is heteroskedastic.

fitted is:  $\text{mutants killed} \propto \sqrt{\text{coverage}}$ , and was chosen because it is something simple that works reasonably well. Searching for the best fitting exponent, using `nls` (the `betareg` package does not support fitting non-linear models), shows that 0.44 is a better fit than 0.5 for this sample.

Not a lot of difference in this case (or this one REXAMPLE[reuse\_and\_prod.R])... Need a better example...

The Beta regression model summary output includes extra information, as follows:

```
Call:
betareg(formula = y_measure ~ I(x_measure^0.5))

Standardized weighted residuals 2:
    Min     1Q   Median     3Q    Max 
-2.6881 -0.6403 -0.1279  0.6399  3.1829 

Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.5460    0.1891 -13.46   <2e-16 ***  
I(x_measure^0.5) 4.7093    0.3502  13.45   <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Type of estimator: ML (maximum likelihood)
Log-likelihood: 80.37 on 3 Df
Pseudo R-squared: 0.6323
Number of iterations: 13 (BFGS) + 1 (Fisher scoring)
```

The  $\phi$  variable is the second coefficient of the fitted Beta equation,  $B(\mu, \phi)$ .

Calls to `predict` return the expected value of the response variable,  $E(y) = \mu$ . The standard deviation in the expected value is given by the following equation (setting `se.fit` does not cause `predict` to return standard error information when passed a Beta regression model):

$$SD(y) = \sqrt{\frac{\mu(1-\mu)}{1+\phi}}$$

The default link function used by the `betareg` function is `logit`, the same default link function used by `glm` when passed the `family=binomial` argument.

$$\log \frac{p}{q} = \alpha + \beta x$$

$p$  proportion of successes,  $q$  proportion of failures.

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

### 11.3.7 Relational responses

List of relational comparisons... an experiment<sup>700</sup>

## 11.4 Multiple explanatory variables

Linear regression can be used to build models containing more than one single explanatory variable; *multiple regression* is the name given to modeling using more than one explanatory variable (the term *bivariate regression* is sometimes applied to the single explanatory variable+response variable case). In theory there is no limit on the number of explanatory variables, but in practice available processing resources and by the need to hold data in storage set an upper bound.

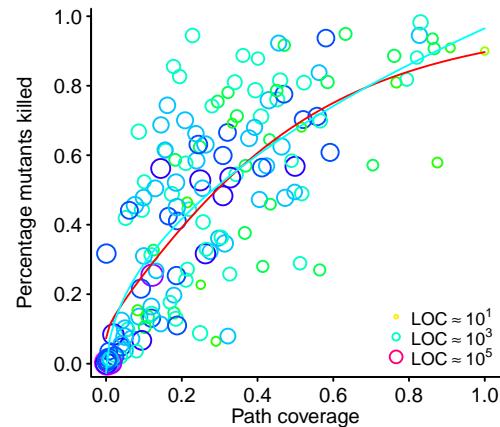


Figure 11.38: Percentage of mutants killed at various percentage of path coverage for 300 or so Java projects; fitted Beta (red) and `glm` (blue) regression models. Data from Gopinath et al.<sup>521</sup> code

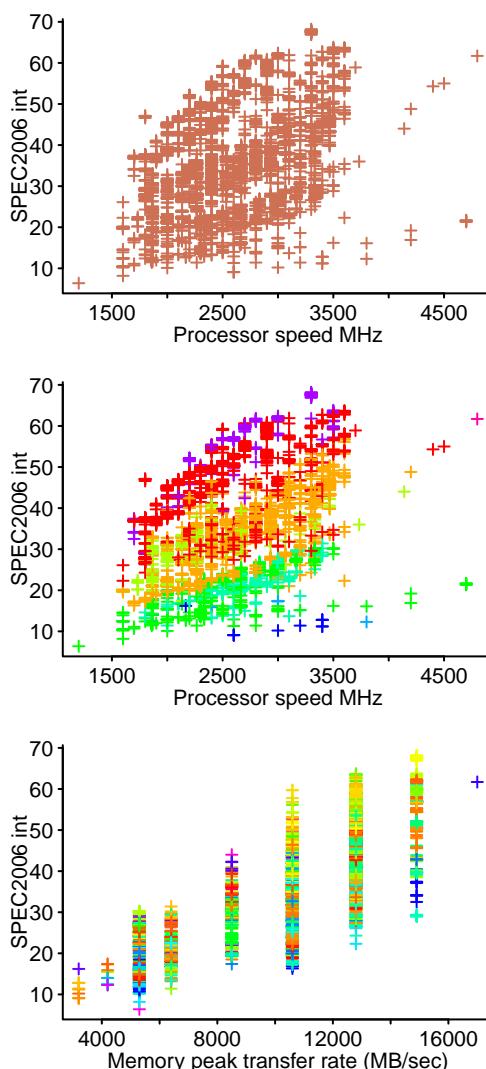


Figure 11.39: SPECint 2006 performance results for processors running at various clock rates, memory chip frequencies and processor family. Data from SPEC.<sup>1282</sup> code

Visualizing data is much more difficult when there are multiple explanatory variables. Chapter 8 contains many examples for visualizing two variables and the general approach is to break down multiple regression visualization into pairs of variables.

System performance is affected by many factors and Figure 11.39 shows SPECint 2006 results for processors running at various frequencies (upper), color coded by memory chip frequency (center) and name of processor family (lower).

The SPECint results include 36 columns of information relating to the benchmarked system. Which of these columns contains information that can be used to succinctly model the performance of a system and what equation best describes the form of their contribution?

The R formula notation provides a means of specifying all columns in the data frame as explanatory variables, except the one specified as the response variable; the dot token is used as follows:

```
spec_mod=glm(Result ~ ., data=cint)
```

Given enough cpu power and memory, it can be more productive to start by considering all explanatory variables, removing underperforming variables, rather than starting with the explanatory variable believed to be the most important and then adding more variables.

The stepAIC function in the MASS package automates the process of removing underperforming explanatory variables from an existing model to create a model having a minimum AIC (the step function in the base system is a rather minimal implementation).<sup>xix</sup>

When some domain knowledge is available (e.g., performance usually correlates with clock rate and is not usually affected by date of execution), experimenting by building models containing those explanatory variables considered to be most likely to have a large impact on the response variable can help refine understanding of the impact of various subcomponent characteristics on overall performance.

For this SPEC dataset there is so much detail recorded in the Processor column of the Spec results, that each entry is often unique; making it possible to create an almost perfect, but completely uninformative, model using just this one explanatory variable.

The following model explains 80% of the variance in the Result values:

```
spec_mod=glm(Result ~ Processor.MHz+mem_rate+mem_freq, data=cint)
```

where: Processor.MHz is the processor clock rate, mem\_rate the peak memory transfer rate and mem\_freq the frequency at which memory is clocked.

The + binary operator, in the above formula, specifies that explanatory variables are added together. The summary output shows that the equation fitted by glm is:

$$Result = -2.4 \cdot 10^1 + Processor.MHz 7.3 \cdot 10^{-3} + mem\_rate 2.5 \cdot 10^{-3} + mem\_freq 1.0 \cdot 10^{-2}$$

Is a linear combination of each explanatory variable the best model? With a single variable, it is easy to visually compare model predictions against measured values and a method of visualizing information for each explanatory variable in a fitted model is required.

The crPlot function, in the car package, produces a *component+residual* plot (also known as a *partial-residual* plot); the y-axis contains the predicted value plus the residual, the x-axis the value of the explanatory variable.

```
library("car")
```

```
spec_mod=glm(Result ~ Processor.MHz+mem_rate+mem_freq, data=cint)

crPlot(spec_mod, variable="Processor.MHz", col=point_col)
crPlot(spec_mod, variable="mem_freq", col=point_col)
crPlot(spec_mod, variable="mem_rate", col=point_col)
```

Figure 11.40 shows the component+residual plots produced using the code above (the red dotted line is derived from the fitted model and the green line a loess fit). If the form of an explanatory variable in the formula used to fit a model is close to reality, the two lines will

<sup>xix</sup> This fishing expedition approach to model building requires that p-values be suitably reduced, e.g., using a Bonferroni corrected value.

be intertwined. For the SPEC model there is obvious curvature for two variables and perhaps some for a third.

Experience of computer behavior suggests that performance will not increase forever, as clock rates are increased. Adding quadratic forms of the explanatory variables to the model is an obvious modification to try in a linear model (an exponential is more realistic in that its value converges to a limit, but this form of modeling requires the use of non-linear regression, which is covered later).

Adding quadratic terms to the model shows that two of the three explanatory variables make worthwhile contributions; see Figure 11.41. The updated model explains another 4% of the variance; perhaps the techniques discussed below can produce further improvements...

Some systems contained error correcting memory, which might be expected to slightly reduce performance. An existing model can be updated to include, or remove, explanatory variables using the update function. The following code adds the variable ecc to the previously built model, spec\_mod:

```
ecc_spec_mod=update(spec_mod, . ~ . + ecc)
```

The advantage of using update is a reduction in the system resources needed to build the model, compared with starting from the beginning again.

The summary output shows that systems containing error correcting memory have slightly better performance. Before jumping to the conclusion that adding error correction improves system performance, it is worth noting that this kind of memory tends to be used in expensive systems, where it is likely that money has been spent to improve performance and reliability.

The cpu and memory frequency are decided by information that is not included in the SPEC results, the intended price point a computing system is designed to be sold at and the trade-off in the cost/performance of the components needed to build it....

How much does each explanatory variable contribute to a fitted model? Ways in which individual contributions can be measured include:

- the amount of variance, in the response variable, explained by an explanatory variable.

The calc.relimp function in the relaimpo package calculates the contribution made by each explanatory variable to the variance explained by the fitted model,

- the impact each explanatory variable can have on the range of values taken by the response variable (with all other explanatory variables maintaining a fixed value).

For very simple models, <sup>xx</sup> one way of calculating the maximum impact on the value of the response variable is by multiplying the minimum/maximum value taken by the explanatory variable by the corresponding coefficient in the fitted model. For instance, `range(cint$Processor.MHz)*7.3*10^-3` evaluates to 11.68 35.04, a difference of 23.36.

The visreg function, in the visreg package, produces a visual representation of the impact of each explanatory variable on the response variable.

Nomograms are a visual technique for calculating the value of a response variable when each explanatory variable has a particular value: the DynNom function in the DynNom package supports interactive exploration of model behavior in a web browser.

Normalising values prior to building a model is sometimes suggested (using the scale function); the relative values of the model coefficients can then be directly compared. This technique only works when all explanatory variable values are drawn from a Normal distribution.

The relaimpo package supports a variety of functions<sup>555</sup> for calculating the relative contribution made to a model by each explanatory variable. For instance, the calc.relimp function can calculate: first, the variance explained by a model containing one variable, last, the variance explained when a variable is added to a model that already contains the other variables, betasq, the standardized coefficients of the model (i.e., one fitted after normalising the data; effectively a metric for the contribution of each explanatory variable to the value of the response variable), and lmg (named after the initials of its creators), the variance explained by each variable; the boot.relimp function returns confidence intervals for these values.

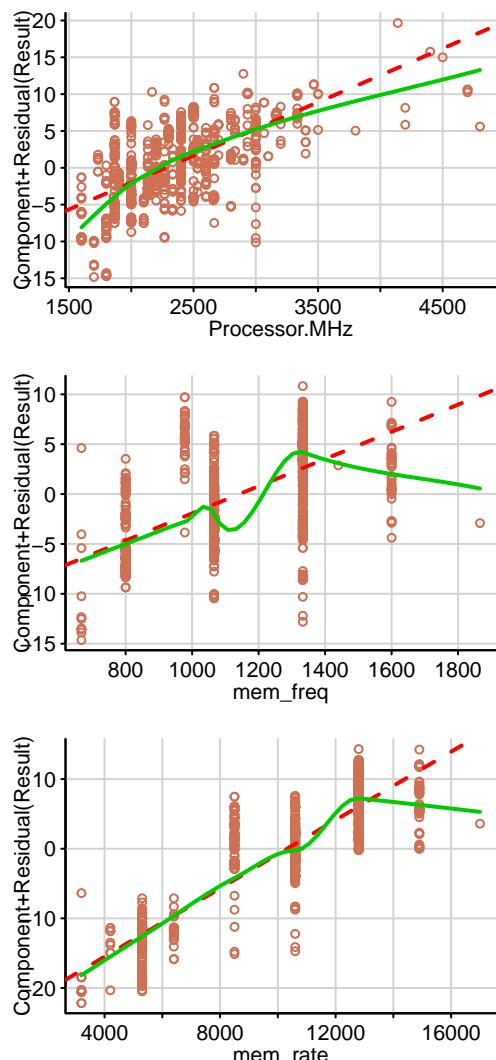


Figure 11.40: Component+residual plots for three explanatory variables in a fitted SPECint model. [code](#)

<sup>xx</sup> Those that are linear in the explanatory variable and no interactions between variables.

```
library("relaimpo")

spec_mod=glm(Result ~ Processor.MHz+I(Processor.MHz^2)+mem_rate
            + I(mem_rate^2)+mem_freq, data=cint)

# How much does each explanatory variable contribute?
calc.relimp(spec_mod, type = c("first", "last", "betasq", "lmg"))
```

In the following output from `calc.relimp`, based on a model built using the SPECint data: [code](#)

```
Response variable: Result
Total response variance: 81.5614
Analysis based on 1346 observations

5 Regressors:
Processor.MHz I(Processor.MHz^2) mem_rate I(mem_rate^2) mem_freq
Proportion of variance explained by model: 83.77%
Metrics are not normalized (rela=FALSE).
```

Relative importance metrics:

	lmg	last	first	betasq
Processor.MHz	0.06188975	0.017806784	0.04608568	0.6888167
I(Processor.MHz^2)	0.04555774	0.005697759	0.02962054	0.2201344
mem_rate	0.29379918	0.028909460	0.55553992	2.0853323
I(mem_rate^2)	0.29050403	0.006362584	0.58253084	0.4768264
mem_freq	0.14598416	0.067530522	0.28997162	0.1258352

Average coefficients for different model sizes:

	1X	2Xs	3Xs	4Xs
Processor.MHz	4.308098e-03	1.289747e-02	1.567592e-02	1.823108e-02
I(Processor.MHz^2)	6.559513e-07	-4.893878e-07	-9.880091e-07	-1.727687e-06
mem_rate	2.343231e-03	1.561771e-03	2.235992e-03	3.303496e-03
I(mem_rate^2)	1.280338e-07	1.488238e-07	8.108289e-08	-1.285720e-08
mem_freq	2.429426e-02	2.012476e-02	1.557607e-02	1.456773e-02

	5Xs
Processor.MHz	1.665538e-02
I(Processor.MHz^2)	-1.788212e-06
mem_rate	4.539887e-03
I(mem_rate^2)	-1.158365e-07
mem_freq	1.600394e-02

the second set of columns, under the line starting Average coefficients, lists the model coefficients for each explanatory variable, if that variable were to appear in a model containing X variables (values are averaged over all combinations of other variables). The values in the last column (5Xs in this case) are the same as those produced by `summary`.

How do changes in the value of each explanatory variable effect the value of the response variable (when the other variables remaining constant)? Figure 11.41 shows the individual contribution made by each explanatory variable to the value of the response variable (along with confidence intervals), when the other variables are held constant, for the following model of SPECint performance:

?

```
library("visreg")

spec_mod=glm(Result ~ Processor.MHz + I(Processor.MHz^2)+mem_freq
            +mem_rate+I(mem_rate^2), data=cint)

visreg(spec_mod)
```

Sometimes including an explanatory that has no correlation with the response variable improves the performance of a model; why does this happen? An explanatory variable may correlate with the residual of a model, so adding this new variable has the effect of improving a model by reducing its residual.

Figure 11.41: Individual contribution of each explanatory variable to the response variable in a quadratic model of SPECint performance. [code](#)

### 11.4.1 Interaction between variables

In the models built so far, each explanatory variable has been independent of the others. The `glm` function and many other regression modeling functions provide mechanisms for specifying interactions between explanatory variables, using binary operators in the formula, such as `:`, `*` and `^`.

Operator	Effect
<code>+</code>	causes both of its operands to be included in the equation.
<code>:</code>	denotes an interaction between its operands, e.g., <code>a:b</code> or <code>a:b:c</code> .
<code>*</code>	denotes all possible combinations of <code>+</code> and <code>:</code> operators, e.g., <code>a*b</code> is equivalent to <code>a+b+a:b</code> .
<code>^</code>	denotes all interactions to a specific degree, e.g., <code>(a+b+c)^2</code> is equivalent to <code>a+b+c+a:b+a:c+b:c</code> .
<code>.</code>	denotes all variables in the data-frame specified in the <code>data</code> argument except the response variable.
<code>-</code>	specifies that the right operand is removed from the equation, e.g., <code>a*b-a</code> is equivalent to <code>b+a:b</code> .
<code>-1</code>	specifies that an intercept is not to be fitted (many regression fitting functions implicitly include an intercept).
<code>I()</code>	"as-is", any operators in the enclosed expression are not treated as formula operators, their behavior that which occurs outside of a formula.

Table 11.2: Symbols that can be used within a formula to express relationships between explanatory variables.

As with all data analysis, the choice of interactions between explanatory variables should be driven by an understanding of the problem domain. When there is a great deal of uncertainty about which interactions are significant, it is often easiest to start by specifying all pairs of interactions between variables (or triple interactions if there are not too many variables) and then to simplify, either automatically using `stepAIC` or through manual inspection of `summary` output of the fitted models.

Stepwise regression techniques, such as that provided by `stepAIC`, can return models that suffer from a variety of problems, such as overfitting. There are techniques available to help avoid these problems; the `train` function in the `caret` package supports some of these techniques.

The `gelmulti` package automates the process of finding an optimal, in a sense specified by the user (e.g., minimise AIC or some other measure), explanatory variable interaction; a list of variables is specified and the function permutes through the possibilities, e.g., `gelmulti("y", c("a", "b", "c", "d"), data=some_data)`.

A study by Moløkken-Østvold and Furulund<sup>967</sup> investigated the impact of daily communication between customer and contractor on the accuracy of effort estimates, for 18 software projects. Figure 11.42 shows estimated vs. actual effort broken down by communication frequency (i.e., daily or not daily), along with individually fitted straight lines.

It is possible to build one regression model that simultaneously fits both straight lines to this data; the following code shows one possibility:

```
sim_mod=glm(Actual ~ Estimated+Estimated:Communication, data=sim)
```

The fitted equation in this case is:

$$\begin{aligned} \text{Actual} &= -270.1 + 1.18\text{Estimated} + 0.51\text{Estimated} \cdot D \\ &= -270.1 + (1.18 + 0.51D)\text{Estimated} \end{aligned}$$

where: `Actual` is the actual and `Estimated` the estimated effort, and `D` has one of two values:

$$D = \begin{cases} 1 & \text{daily communication} \\ 0 & \text{not daily communication} \end{cases}$$

Is this formula the best fit possible using the available data? The formula used in this model was selected by your author because of a belief that the benefit of communication will increase as project size increases.

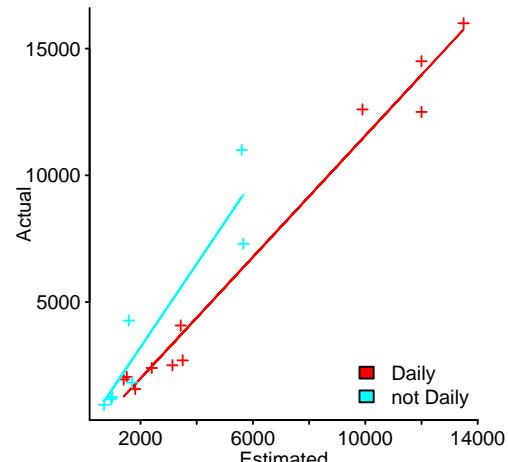


Figure 11.42: Estimated and actual effort broken down by communication frequency, along with individually fitted straight lines. Data from Moløkken-Østvold et al.<sup>967</sup> code

There are six data points for each of the 18 projects, computationally small enough for the brute force approach of examining all possible models; but with only 18 projects, some formula possibilities cannot be fitted because they contain more variables than available data points (a unique solution requires fewer variables than data points).

The formula in the following code fits four explanatory variables individually, plus each variable paired with every other variable (one at a time). `stepAIC` is used as a quick way of removing explanatory variables that are not paying their way (automatic model selection is fraught with problems, with perhaps the largest being that it causes users to stop thinking):

```
sim_mod=glm(Actual ~ (Estimated+Communication+Contract+Complexity)^2, data=sim)
min_sim=stepAIC(sim_mod)
summary(min_sim)
```

This book's primary aim is to build models as a means of developing understanding, and minimising AIC is often a useful step along the way. Another possible way of removing low impact variables from a model is to consider the p-value of each fitted component.

The `summary` output for ordinal and nominal explanatory variables lists p-values for each value that these variables take in the data. The `Anova` function (in the `car` package) lists p-values at the variable level and its output for the above model is: `code`

`Analysis of Deviance Table (Type II tests)`

Response: Actual	LR	Chisq	Df	Pr(>Chisq)							
Estimated	45.879	1	1.258e-11	***							
Communication	17.272	1	3.240e-05	***							
Contract	6.767	3	0.07971	.							
Complexity	1.543	1	0.21423								
Estimated:Communication	2.546	1	0.11060								
Communication:Contract	5.020	3	0.17034								
Contract:Complexity	3.197	2	0.20224								
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

The variable `Complexity` has the highest p-value and repeatedly removing the component having the highest p-value, for successively smaller models leads to the following model:

```
sim_mod=glm(Actual ~ Estimated+Communication+Communication:Contract,
            data=sim)
```

which fits the equation:

$$\begin{aligned} \text{Actual} = & -274.8 + 1.21\text{Estimated} + 2625 \cdot !D + \\ & C_{fp}(1862 \cdot !D - 197.6 \cdot D) + \\ & C_{tp}(-2270 \cdot !D - 462.2 \cdot D) + \\ & C_{ot}(-2298 \cdot !D - 234.3 \cdot D) \end{aligned}$$

where the new variables are:  $C_{fp}$  is a fixed price contract,  $C_{tp}$  is a target price contract and  $C_{ot}$  other kind of contract.

$$C_{fp} = \begin{cases} 1 & \text{fixed price contract} \\ 0 & \text{not fixed price contract} \end{cases} \quad C_{tp} = \begin{cases} 1 & \text{target price contract} \\ 0 & \text{not target price contract} \end{cases} \quad C_{ot} = \begin{cases} 1 & \text{other contract} \\ 0 & \text{not other contract} \end{cases}$$

This model explains more of the variance in the data than the first model built, it also has a slightly smaller AIC. While it makes use of extra information (i.e., the kind of contract), a more noticeable difference is that `Communication` has a constant effect (i.e., it does not increase with estimated size); the case of fixed price contracts with no daily communication cries out for attention.

Following the numbers has produced a model which better fits the data, but does not fit expectation (which may, of course, be wrong).

### 11.4.2 Correlated explanatory variables

The mathematics behind many of the approaches used to build linear regression models assumes that explanatory variables are independent of each other. If a linear relationship

exists between one or more pairs of explanatory variables (i.e., a relationship of the form:  $PV_1 = a + b \times PV_2$ , where  $PV_1$  and  $PV_2$  are explanatory variables, and  $a$  is any constant and  $b$  is a non-zero constant), then this needs to be taken into account by the model building technique used.<sup>xxi</sup>

*Multicollinearity* is said to occur when a linear relationship exists between two or more explanatory variables, the term *colinearity* is often used when only two variables are involved.

Figure 11.43 illustrates how the variance in  $Y$  explained by combining  $X_1$  and  $X_2$  may be less than the sum of the variance explained by each individually, because the two variables are not independent; there is a shared contribution.

The impact of multicollinearity is to increase the standard error in the calculated value of the coefficients of the fitted model (i.e., the  $\beta_n$ ), potentially resulting in a model that is not considered acceptable or being unreliable (in the sense that small changes in the data result in large changes in the coefficients of the fitted model). The increased uncertainty in some variables will make it more difficult to isolate the effects of individual explanatory variables and will increase the width of the confidence intervals for the predicted values of the response variable.

The *Variance Inflation Factor* (VIF) is a measure of the uncertainty created by the presence of multicollinearity. The impact of VIF is the same as reducing the sample size (when no multicollinearity is present, VIF has a value of one):

$$\epsilon_{\text{standard}} \propto \sqrt{\frac{\text{VIF}}{\text{observations}}}$$

When is a VIF value too large? A large VIF is more likely to be acceptable when there are many observations, compared to when there are few, e.g., the standard error is proportionally the same for 10,000 observations having a VIF of 400 and for 100 observations having a VIF of 4.

Suggested maximum VIF values do appear in print, e.g., 5 or 10 are sometimes suggested. As always, think about what the VIF value means in the context of how the results will be used; pick a value that makes sense given the sample size, the error in the measurements and the level of error that is acceptable in the business context.

The car and rms packages support a `vif` function, taking the model returned by a call to, for instance, `glm` and returning the VIF for each explanatory variable.

A study by Kroah-Hartman<sup>548</sup> investigated the amount of change in the Linux kernel source code occurring between each release. Figure 11.44 shows the number of lines added, modified and removed, plus overall growth, number of files and total number of lines at each initial release of the Linux kernel from version 2.6.0 to 3.9 (two outliers have been excluded).

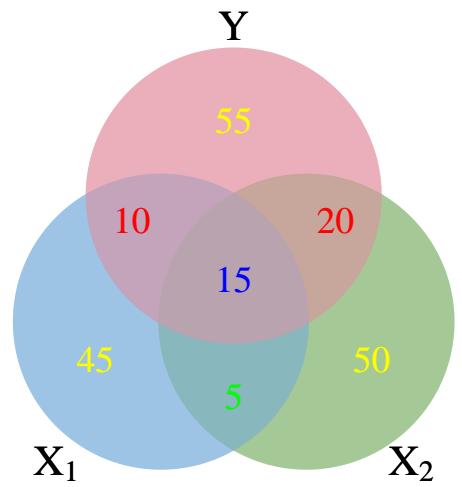


Figure 11.43: Illustration of the shared and non-shared contributions made by two explanatory variables to the response variable  $Y$ . [code](#)

<sup>xxi</sup>It is ok for a nonlinear relationship to exist e.g.,  $PV_1 = a + b \times PV_2^2$ .

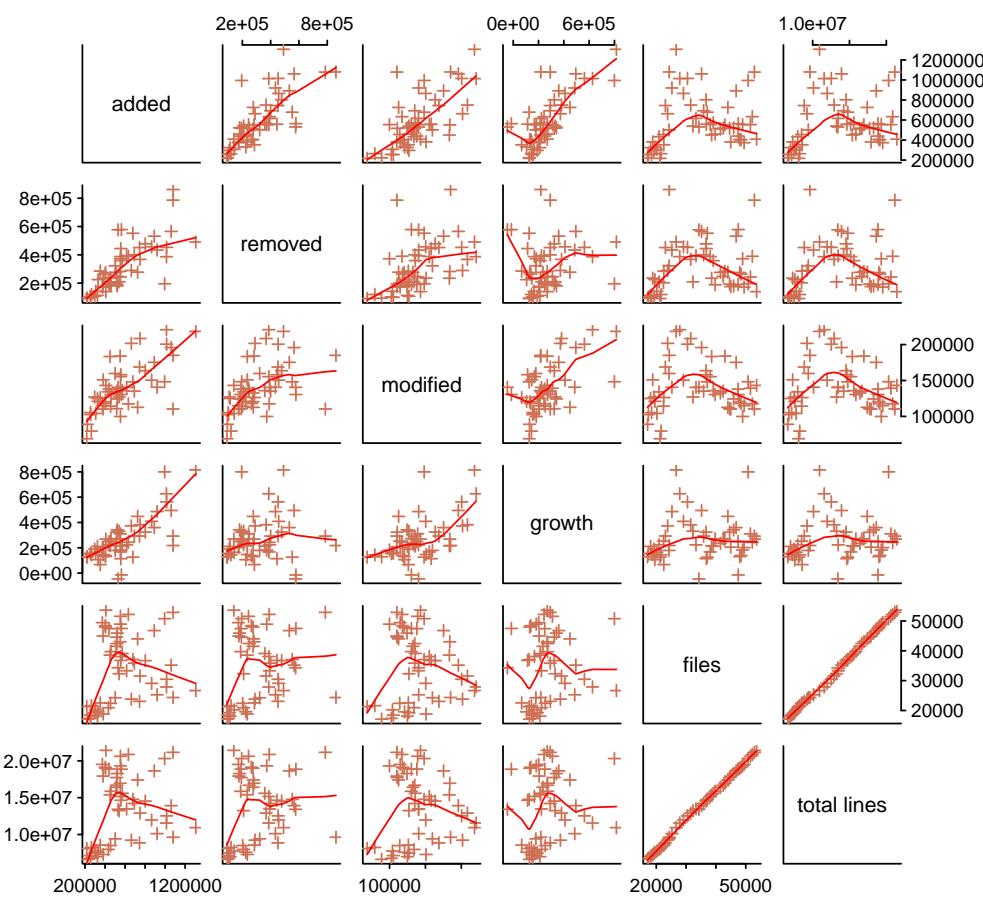


Figure 11.44: pairs plot of lines added/modified/removed, growth and number of files and total lines in versions 2.6.0 through 3.9 of the Linux kernel. Data from Kroah-Hartman.<sup>548</sup> [code](#)

Building a model of the growth of the Linux kernel is complicated by the potentially large amount of correlation between some of the measured variables, including:

- the growth in, lines of code, between releases is the difference between lines added and lines removed; these three variables are perfectly correlated in that knowing two of them enables the third to be calculated,
- lines added appears strongly correlated with lines removed. Perhaps existing functionality is being rewritten, rather than being completely new,
- the decision about whether a line has been modified or removed/added is made algorithmically (rather than asking the developer who made the change). The amount of misclassified lines is not known,
- system level measurements are also correlated, e.g., number of files and total lines of code.

Modeling the number of modified lines, using the Kroah-Hartman data, finds that both lines added and lines removed individually explain around half of the deviance (61% and 41% respectively). However, combining them both appear in a model does not produce any improvement; the following output from `summary` was obtained by including the argument `correlation=TRUE`. [code](#)

```
Call:
glm(formula = lines.modified ~ lines.added + lines.removed, data = amr_out)

Deviance Residuals:
    Min      1Q  Median      3Q     Max  
-72376 -12049     321   11274   54964  

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.705e+04 8.625e+03 10.093 9.40e-14 ***
lines.added  9.958e-02 2.093e-02  4.759 1.64e-05 ***
lines.removed -1.500e-02 3.117e-02 -0.481    0.632  
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for gaussian family taken to be 639477748)
```

```
Null deviance: 6.2276e+10 on 53 degrees of freedom
Residual deviance: 3.2613e+10 on 51 degrees of freedom
AIC: 1253.1
```

Number of Fisher Scoring iterations: 2

```
Correlation of Coefficients:
      (Intercept) lines.added
lines.added -0.54
lines.removed -0.06     -0.76
```

The correlation between the model coefficients appears at the end of the output and shows a high correlation between `lines.added` and `lines.removed`; `lines.added` is a better predictor of `lines.modified` and has been selected over `lines.removed` (whose p-value is significantly larger than when this variable appeared on its own in a model).

A call to the `vif` produces the following: [code](#)

```
lines.added lines.removed
2.39311    2.39311
```

With only two explanatory variables there is no ambiguity about which variables are involved in a linear relationship, but when more than two variables are involved things are not always so obvious. The correlation output from `summary` can be used to identify related variables; the `alias` function generates just this information when the argument `partial=TRUE` is specified.

Approaches to dealing with multicollinearity having an undesirable impact on the fitted model include:

- removing one or more of the correlated explanatory variables. The choice of which explanatory variables to remove might be driven by:
  - the cost of collecting information on a variable,
  - a VIF driven approach. The process builds a model using the current set of explanatory variables, removes the explanatory variable with the largest VIF (removing one variable effects the VIF of those that remain and may reduce the VIF of other variables to an acceptable level) and repeats until all explanatory variables have what is considered to be an acceptable VIF,
- combining the strongly correlated variables in a way that makes use of all the information they contain.

The disadvantages of excluding explanatory variables from a model include:

- ignoring potentially useful information present in the excluded variable,
- the resulting model may give a false impression about which explanatory variables are important, i.e., readers will assume that the variables appearing in the model are the only important ones, unless information about the excluded variables is given,
- it provides a means for the data analyst to select a model that favours the hypothesis they want to promote (by allowing them to select which explanatory variables appear in the model).

The SPEC power benchmark<sup>[1283](#)</sup> is designed to measure single and multi-node server power consumption while executing a known load. The results contain 515 measurements of six system hardware characteristics, such as number of chips, number of cores and total memory, as well as average power consumption at various load factors.

A model of average power consumption, at 100% load, containing a linear combination of all explanatory variables, shows very high multicollinearity for the number of chips (its VIF is 27.5 and several other variables have a high VIF; see [hardware/SPECpower.R](#)). Removing

this variable reduces the VIF of the remaining variables, but the AIC drops from 6798.7 to 7182.1. Whether this decrease in model performance is an important issue depends on the reason for building the model, i.e., prediction or understanding. Do the values of the model coefficients, after removing this variable, provide more insight than the coefficient values of the original model? These kinds of questions can only be answered by a person having detailed domain knowledge. This example illustrates that removing a variable solves one problem and raises others.

A study of fault prediction by Nagappan, Zeller, Zimmermann, Herzig and Murphy<sup>988</sup> produced data containing six explanatory variables having an exact linear relationship with other explanatory variables. The `glm` function detects the existence of this relationship and excludes the offending explanatory variables from the model (the value returned for their fitted coefficients is NA); see [regression/change-burst-sum.R](#).

Two explanatory variables having an exact linear relationship will have a correlation of  $\pm 1$ , as a call to `alias` will show.

### 11.4.3 Penalized regression

*Penalized regression* handles multicollinearity by using a technique that automatically selects how much each explanatory variable should contribute to the model; explanatory variables are penalized based on their relative contribution to the model.

The traditional technique for fitting a regression model involves minimising some measure of the error, where the error is defined to be the difference between actual and predicted values, e.g., the sum of squared error, and the equation is:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Penalised regression modifies this equation to include a penalty (the  $\lambda$  in the equation below) for the  $P$  coefficients in the model ( $\beta$  in the equation below).

$$SSE_{enet} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^P |\beta_j| + \lambda_2 \sum_{j=1}^P \beta_j^2$$

This technique of using both first- and second-order penalties is known as the *elastic net*. When only the first order term,  $\lambda_1$ , is used it is known as the least absolute shrinkage and selection operator method (*lasso*). When only the second order term,  $\lambda_2$ , is used it is known as *ridge regression*.

In theory the penalization penalties, the values of  $\lambda_1$  and  $\lambda_2$ , are chosen by the user. In practice packages provide a function that automatically finds values (using bootstrap) that minimise the error.

The lasso tends to pick one from each set of correlated variables and ignore the rest (by setting the corresponding  $\beta$ s to zero). Ridge regression has the effect of causing the coefficients,  $\beta$ , of the corresponding correlated variables to converge to a common value, i.e., the coefficient chosen for  $k$  perfectly correlated variables is  $\frac{1}{k}$ th the size chosen had just one been used.

The calculation of mean squared error adds contributions from both variance and bias. The default regression modeling techniques are unbiased (i.e., attempt to minimise bias). It is possible to build models with lower MSE by trading off bias for variance; one consequence of correlation between variables is that the variance is high...

The `penalized` package...

now if we had some data where it makes a big difference (see [hardware/SPECpower.R](#) for a case where this makes a small difference)...

## 11.5 Non-linear regression

The term linear is applied to the regression models built so far because the coefficients of the model (e.g.,  $\beta_1$  in the equation at the start of this chapter) are linear (the form of the explanatory variables is irrelevant). In a non-linear regression model one or more of the coefficients appear in a context that creates a non-linear equation, as in the following:

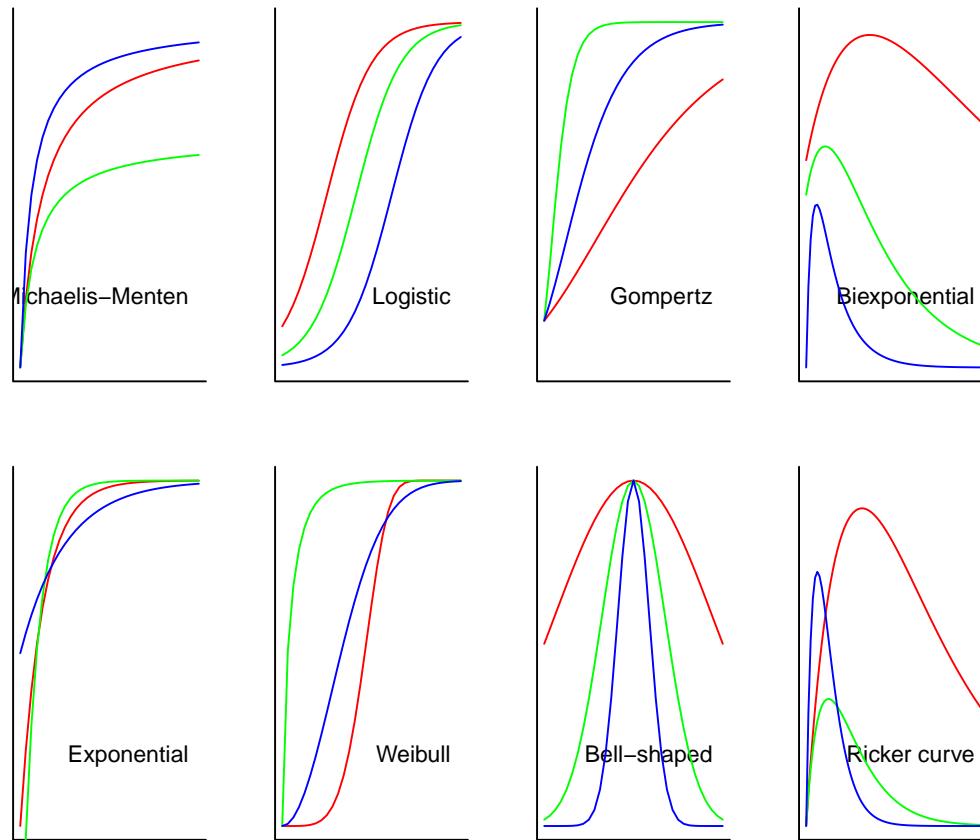
$$y = \alpha + \beta_1 x^{\beta_2} + \epsilon$$

Table 11.3 lists some commonly occurring non-linear equations and Figure 11.45 illustrates some example instances of these equations.

The `nls` function (Nonlinear Least Squares) is part of the base system and can be used to build non-linear regression models. This function requires that the response variable error distribution be Normal (the default behavior used by `glm`). The `gnm` package (Generalized nonlinear models) contains support for other forms of error distribution.

Shape	Name	Equation
Asymptotic growth to a limit	Michaelis-Menten	$y = \frac{ax}{1+bx}$
Asymptotic growth to a limit	Exponential	$y = a(1 - be^{-bx})$
S-Shaped	Logistic	$y = a + \frac{b-a}{1+e^{(c-x)/d}}$
S-Shaped	Weibull	$y = a - be^{-cx^d}$
S-Shaped	Gompertz	$y = ae^{be^{-cx}}$
Humped	Bell-shaped	$y = ae^{- bx ^2}$
Humped	Biexponential	$y = ae^{-bx} - ce^{-dx}$
Humped	Ricker curve	$y = axe^{-bx}$

Table 11.3: Some commonly encountered non-linear equations, see Figure 11.45.



From the practical point of view there are several big differences between using `glm` and using `nls`, including:

- `nls` may fail to fit a model; the techniques used to find the coefficients of a non-linear model are not guaranteed to converge,
- `nls` may return a fitted model that differs from the actual solution; the techniques used to find the coefficients of a non-linear model may become stuck in a local minimum that is good enough and not find a better solution,
- `nls` often requires users to provide an estimate for the initial value for each model coefficient, that is close to the final values (the `start` argument),

Figure 11.45: Example plots of functions listed in Table 11.3. These equations can be inverted, so they start high and go down. [code](#)

- the names of the model coefficients being estimated have to explicitly appear in the formula,
- the operators appearing in the expression to the right of  $\sim$  have their usual arithmetic interpretation, i.e., the behaviors listed in Table 11.2 do not apply.

The biggest problem with fitting non-linear regression models is finding a combination of starting values that enable `nls` to converge to a fitted model. Possible techniques include:

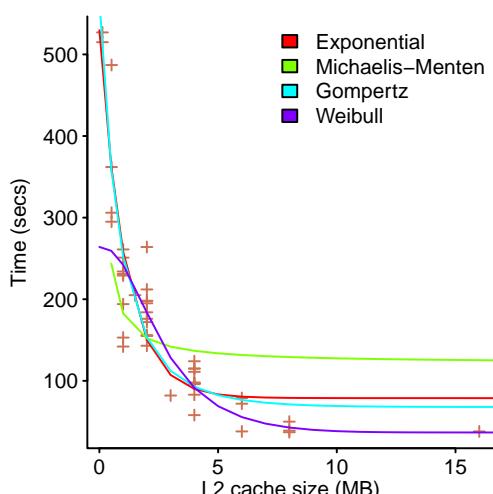
- using a "self-start" function, if available (e.g., `SSlogis` for Logistic models), these attempt to find good starting values to feed into `nls`. These function, in turn, require starting values, but at least there is a known method for calculating them,
- fitting a linear model that is close enough to the non-linear model and using the coefficients of the fitted linear model as starting values,
- using the argument `trace=TRUE`, which outputs the list of model coefficients being tried internally, as a source of ideas to try,
- picking a few points in the plotted data that a fitted line is likely to pass through and calculate values that would result in an equation used passing close to these points.

A study by Hazelhurst<sup>593</sup> measured the performance of various systems running a computational biology program. Figure 11.46 shows four non-linear equations fitted to one processor characteristic (L2 cache size). The calls to `nls` are as follows:<sup>xxii</sup>

```
b_mod=nls(T1 ~ c+a*exp(b*L2), data=bench,
           start=list(a=300, b=-0.1, c=60))

mm_mod=nls(T1 ~ (1+b*L2)/(a*L2), data=bench,
            start=list(b=3, a=0.004))

gm_mod=nls(T1 ~ a/exp(b*exp(-c*L2)), data=bench,
            start=list(a=80, b=-1, c=0.1), trace=FALSE)
Asym = 0.0125
Drop = 0.002
lrc = -1.0
pwr = 2.5
# 1/SSweibull does not have the desired effect, so have to invert
# the response.
getInitial(1/T1 ~ SSweibull(L2, Asym, Drop, lrc, pwr), data=bench)
wb_mod=nls(1/T1 ~ SSweibull(L2, Asym, Drop, lrc, pwr), data=bench)
```



At the start of this chapter, Figure 11.7, various linear models were fitted to the growth of Linux; polynomials containing integer powers were used, perhaps the data is better fitted by a polynomial containing non-integer powers. The following call to `nls` attempts to fit such an equation using start values extracted from the quadratic model fitted earlier:

```
m1=nls(LOC ~ a+b*Number_days+Number_days^c, data=h2,
       start=list(a=3e+05, b=-4e+2, c=2.0))
```

The summary and AIC output is: `code`

```
Formula: LOC ~ a + b * Number_days + Number_days^c
```

Parameters:

	Estimate	Std. Error	t value	Pr(> t )
a	-1.679e+05	2.969e+04	-5.656	2.61e-08 ***
b	7.319e+02	3.463e+01	21.131	< 2e-16 ***
c	1.806e+00	4.616e-03	391.211	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 231800 on 498 degrees of freedom

Number of iterations to convergence: 5

Achieved convergence tolerance: 4.299e-06

```
[1] "AIC = 13805.2100165816"
```

---

<sup>xxii</sup> It is difficult to separate inspiration from suck it and see in this process.

showing that the equation:

$$sloc = (-1.68 \cdot 10^5 \pm 3 \cdot 10^4) + (7.32 \cdot 10^2 \pm 3.5 \cdot 10^1) Number\_days + Number\_days^{1.81 \pm 4.6 \cdot 10^{-3}}$$

is a slightly better fit than a cubic equation (i.e., a lower AIC) and also predicts continuing growth (unlike the cubic equation).

It is possible that further experimentation will find a polynomial model with a lower AIC. However, the purpose of this analysis is to understand what is going on, not to find the equation whose fitted model has the lowest AIC.

A more practical issue to address is the creation of a model that makes what are considered to be more realistic future predictions. The growth in the number of lines of code in the Linux kernel will not continue forever, at some point the number of lines added will be closely matched by the number of lines deleted. One commonly seen growth pattern starts slow and then has a rapid growth period followed by a levelling off converging to an upper limit (i.e., an S-shaped curve). The Logistic equation is S-shaped and is often used to model this pattern of growth; the equation involves four unknowns (fourth row in Table 11.3).

```
# suck it and see...
m3=nls(LOC ~ a+(b-a)/(1+exp((c-Number_days)/d)), data=h2,
       start=list(a=-3e+05, b=4e+6, c=2000, d=800))
# no thinking needed, SSfpl works out of the box for this data :-)
m3=nls(LOC ~ SSfpl(Number_days, a, b, c, d), data=h2)
```

The AIC for the fitted Logistic equation is slightly worse than the cubic polynomial (13,273 vs 13,220), but a lot better than the quadratic fit and predicts a future trend that might be considered more likely to occur.

While the predict function includes parameters to request confidence interval and standard error information, support for both is currently unimplemented. The confint function in the MASS package, when passed a model built using nls, returns the confidence intervals for each of the model coefficients. Alternatively, bootstrapping can be used to find confidence intervals.

Figure 11.47 shows the fitted model predicting a slow down in growth and the maximum being reached at around 10,000 days. Who is to say whether this prediction is more likely to occur, over the specified number of days than the continuing increase predicted by the quadratic model? Given that the one explanatory variable used to fit the models, time, has no direct impact on the production of source it is no surprise that the predictions of future behavior made by the various models vary so wildly.

One technique for getting a rough idea of the accuracy of the future predictions made by a model is to fit models to subranges of the available data, and then check the predictions made against the known data outside the subrange. Figure 11.48 shows logistic equations fitted to various subranges of the data, e.g., all data up to 2900, 3650, 4200 number of days and all days.

The lesson to learn from Figure 11.48 is to be careful what you ask for, if you ask for a logistic equation fitted to the data, you may get one. The fitting process is driven by your expectations (in the form of a formula) and the data it is given.

The processes generating the data fitted by a Logistic equation may only broadly follow this pattern, with independent processes each making separate contributions. A study by Grochowski and Fontana<sup>554</sup> showed that increases in the density of data stored on hard disks could be viewed as a sequence of technologies that each rapidly improved (e.g., magneto-resistive and antiferromagnetically-coupled). Figure 11.49 shows the areal density (think magnetic domains) of various models of hard disk on first entering production. Improvements in each technology can be fitted with its own Logistic equation, as can the overall pattern of performance improvements.

A codebase showing some evidence of having completed its major expansion phase us glibc (i.e., its growth rate has levelled off), the GNU C library; see Figure 11.50. Plugging the fitted model coefficients into the Logistic equation we get:

$$y = -28168 + \frac{1114626 + 28168}{1 + e^{(3652-x)/935}}$$

Since the measurements fitted were made the C Standard's committee, JTC1 SC22/WG14, have started work on revising the current specification; so the model's prediction that glibc will max out at around 1,115,000 lines is unlikely to come true.

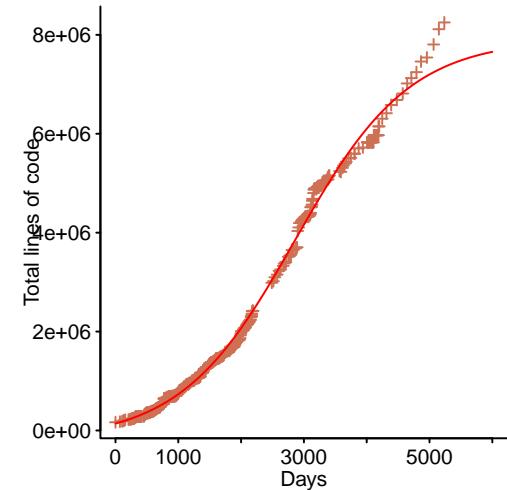


Figure 11.47: A logistic equation fitted to the lines of code in every non-bugfix release of the Linux kernel since version 1.0. Data from Israel et al.<sup>673</sup> [code](#)

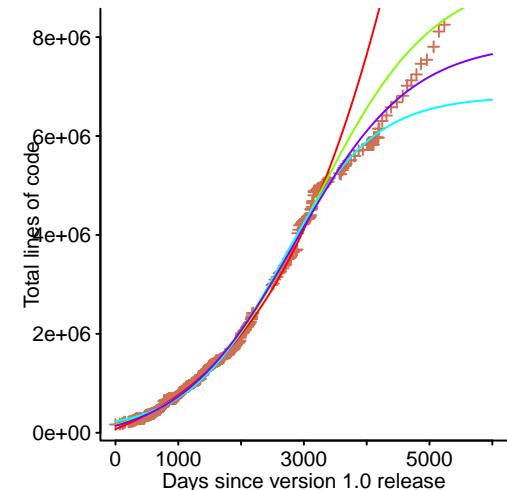


Figure 11.48: Predictions by logistic equations fitted to Linux SLOC data, using subsets of data up to 2900, 3650, 4200 number of days and all days since the release of version 1.0. Data from Israel et al.<sup>673</sup> [code](#)

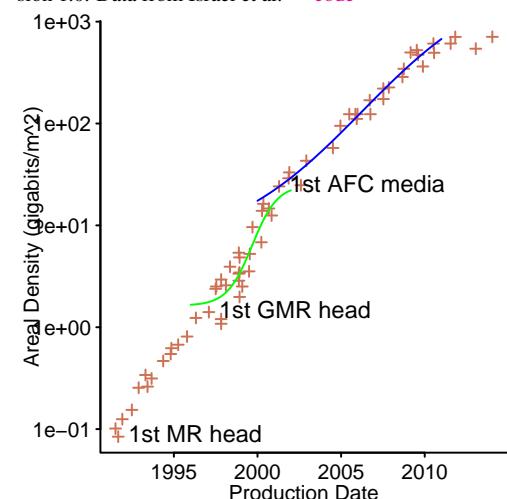


Figure 11.49: Increase in areal density of hard disks entering production over time. Data from Grochowski et al.<sup>554</sup> [code](#)

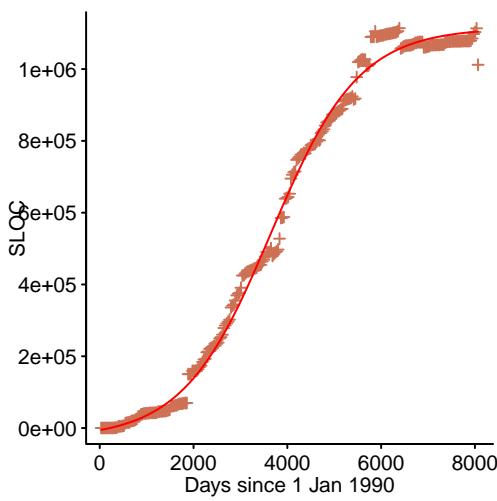


Figure 11.50: Lines of code in the GNU C library against days since 1 January 1990. Data from González-Barahona.<sup>514</sup> [code](#)

A study by Chen, Groce, Fern, Zhang, Wong, Eide and Regehr<sup>244</sup> investigated faults in a C compiler and JavaScript engine by having them process randomly generated programs. Some programs failed to be processed correctly (1,298 in gcc and 2,603 in Mozilla's SpiderMonkey) and many of these failures could be traced back to the same few underlying compiler faults, i.e., some faults were encountered more often than others. Figure 11.51 shows the number of failing programs that could be traced to the same compiler fault, the curved green line is a regression fit (a biexponential, or double exponential); the two straight lines are the exponentials added to create the fit.

The `nls` has a `SSbiexp` starter function, which performs poorly for this data (or, at least, your author could not make it do better).

The sample contains count data, with many very small values, implying a Poisson error distribution. The `gnm` function, in the `gnm` package, has an option to select an error distribution.

The formula notation used by `gnm` is based on function calls,<sup>1371</sup> rather than the binary operators used by `glm` and `nls`. The formula argument in the following call (used to fit the model plotted in Figure 11.51) contains two exponentials (specified using the `instances` function), specified as a constant (the literal 1 is a placeholder for an unknown constant) multiplied (the `Mult` function) by an exponential (the `Exp` function); as with calls to `nls`, starting values are required:

```
library("gnm")

fail_mod=gnm(count ~ instances(Mult(1, Exp(ind)), 2)-1,
             data=wrong_cnt, verbose=FALSE,
             start=c(2000.0, -0.6, 30.0, -0.1),
             family=poisson(link="identity"))
```

A possible reason for why the biexponential is such a good fit is discussed elsewhere, see Figure 6.19.

### 11.5.1 Power laws

Plotting values drawn from a power law distribution using a log scale for both axis, produces a straight line. This straight line characteristic is not unique to power laws and can also be seen in values drawn from other distributions over a wide range of values, e.g., an exponential distribution.<sup>xxiii</sup>

The `powRlaw` package includes functions for fitting and checking whether a power law is likely to be a good fit for a sample.<sup>263</sup>

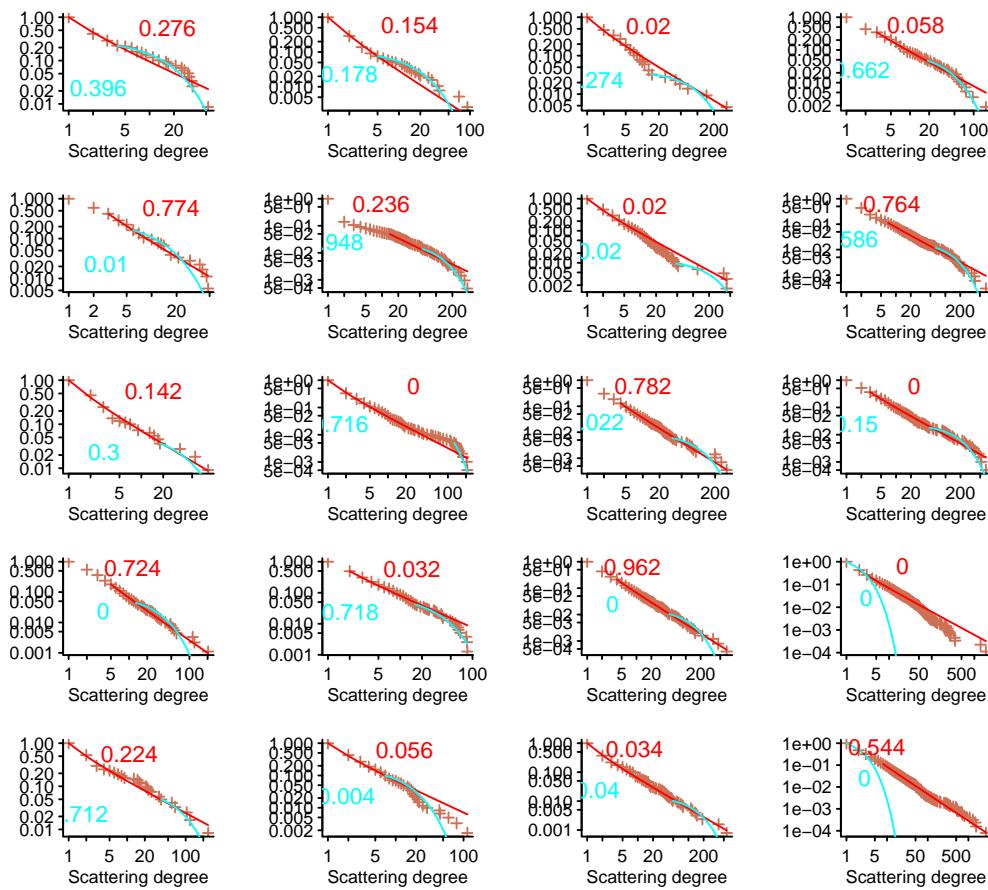
When the model being fitted contains only one explanatory variable having the form of a power law, use of functions from the `powRlaw` package is the recommended approach. However, this package does not support more complicated models and so other functions have to be used when a power law is one of multiple components in a model, e.g., `nls`.

A study by Queiroz, Passos, Valente, Hunsen, Apel and Czarnecki<sup>1134</sup> analysed the conditional compilation directives (e.g., `#ifdef`) used to control the optional features in 20 systems written in C. Researchers in this area use the term *feature constants* to denote macro names used to control the selection of optional features and *scattering degree* to describe the number of `ifdefs` that refer to a given feature constant, e.g., if the macro `SUPPORT_X` appears in two `ifdefs`, it has a scattering degree of two.

Figure 11.52 shows the total number of feature constants (y-axis) having a given scattering degree (x-axis) in these 20 systems. A power law (red) and exponential (blue) is fitted to the data; the numbers are the p-values for the fit (higher is better, i.e., fail to reject the hypothesis). This is a fishing expedition involving 20 systems and a power law is suggested by the visual form of the plotted data and with multiple tests it is necessary to take into account the increased likelihood of a chance match.

If 0.05 is taken as the p-value cutoff, below which the distribution hypothesis is rejected for one test, then  $(1 - 0.95^{20}) \rightarrow 0.64$  is the cutoff when 20 tests are involved. Some systems have p-values above the cutoff for one of the power law or exponential fitted models and so the chosen distribution is not rejected for these systems.

<sup>xxiii</sup> Papers<sup>862</sup> claiming to have found a power law purely on the basis of a plot showing points scattered roughly along a straight line are surprisingly common.



The poweRlaw package supports discrete and continuous forms of heavy tailed distributions. The scattering degree is an integer value and the code below fits both a discrete power law and exponential to the data (the continuous forms are `conpl` and `conexp` respectively):

The power law equation includes a minimum value of  $x$ , scattering degree in this case, below which it does not hold. The `estimate_xmin` function estimates the value,  $x_{min}$ , that minimises the error between the fitted model and the data. The new function, called by the constructor, sets  $x_{min}$  to the minimum value present in the data. It is common for power laws to fit a subset of the data.

```
library("poweRlaw")

# Fit scattering degree
# displ is the constructor for the discrete power law distribution
pow_mod=displ$new(FS$sd)
exp_mod=disexp$new(FS$sd) # discrete power exponential

# Estimate the lower threshold of the fit
pow_mod$setXmin(estimate_xmin(pow_mod))
exp_mod$setXmin(estimate_xmin(exp_mod))

# Plot sample values
plot(pow_mod, col=point_col, xlab="Scattering degree", ylab="")
lines(pow_mod, col=pal_col[1]) # Plot fitted line
lines(exp_mod, col=pal_col[2])

# Bootstrap to test hypothesis that sample drawn from a power law
bs_p=bootstrap_p(pow_mod, threads=4, no_of_sims=500)
text(40, 0.5, bs_p$p, pos=2, col=pal_col[1]) # Display value
```

Figure 11.52: Power law (red) and exponential (blue) fits to feature macro usage in 20 systems written in C; fail to reject p-value for 20 systems is 0.64. Data from Queiroz et al.<sup>1134</sup> [code](#)

## 11.6 Mixed-effects models

A study by Balaji, McCullough, Gupta and Agarwal<sup>90</sup> measured the power consumption of six different Intel Core i5-540M processors executing the SPEC2000 benchmark at various

clock frequencies. The six processors measured are a sample of the entire population of Intel Core i5-540M processors. The power consumption characteristics of this might be modeled using the combined data from all six processors; the following is the summary output for this model: [code](#)

```
Call:
glm(formula = meanpower ~ frequency, data = power_bench)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5746	-0.1882	0.0413	0.1902	2.2965

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.12594	0.01506	141.2	<2e-16 ***
frequency	1.95248	0.00767	254.6	<2e-16 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1429928)

```
Null deviance: 10692.7 on 9980 degrees of freedom
Residual deviance: 1426.9 on 9979 degrees of freedom
AIC: 8916.2
```

Number of Fisher Scoring iterations: 2

This model does not provide any information about how performance varies between processors. The identity of the processor measured could be included in the model (see [regression/hotpower.proc.R](#)), but this cannot be generalized to the entire population.

Possible techniques for estimating model variability caused by processor differences include:

- build a regression model for each processor and then average these six models in some way, e.g., use the coefficients from the six models to build a regression model that is a model of models,

Electronic circuit theory tells us that processor power consumption is proportional to clock frequency and Figure 11.53 shows the results of fitting a separate straight line to the data for each processor.

- build a *mixed-effects model*. A mixed-effects model<sup>xxiv</sup> might be viewed as a model of models; mathematically it uses a more direct approach that makes more effective use of the available data than the method described above.

In a mixed-model the explanatory variables are classified as either a *fixed-effect* or *random-effect* (sometimes called a *covariate*). Technically the effects are not fixed and are not random<sup>xxv</sup>. One way to think about classifying the two kinds of explanatory variables is to look at the impact they have on the response variable:

- fixed effects influence the mean value of the response variable and are associated with the entire population,
- random effects influence the variance of the response variable and are associated with individual subjects.

Mixed-effects models are used to model measurements of clusters of related subjects and multiple correlated measurements of the same subjects (e.g., before/after measurements of the same subject).

A number of different packages are available for building mixed-effects models, the one primarily used in this book is `lme4`, whose workhorse functions are the `glmer` and `lmer` functions<sup>xxvi</sup>.

<sup>xxiv</sup> Also known as a *hierarchical model*.

<sup>xxv</sup> Some authors point this out and then proceed to use what they consider to be more technically correct terms, this book follows common usage because it is common; these terms crop up as named parameters in functions and appear in output information.

<sup>xxvi</sup> A call to the `glmer` function that uses the default family distribution, i.e., `gaussian`, generates a warning that this usage is deprecated and `lmer` should be used.

The lme4 package extends the formula notation to support the specification of random effects. In the following code:

```
library("lme4")

# Express in Gigahertz (otherwise lmer does not converge)
power_bench$frequency=power_bench$frequency/1000000

p_mod=lmer(meanpower ~ frequency + (1 | processor), data=power_bench)
p_mod=lmer(meanpower ~ frequency + (frequency-1 | processor), data=power_bench)
p_mod=lmer(meanpower ~ frequency + (frequency | processor), data=power_bench)
```

- first call to lmer: frequency is the fixed-effect and  $(1 | \text{processor})$  is the random-effect; the 1 specifies there is variation in the value of the intercept and the source of this variation is the processor variable (i.e., the column having this name in the data frame). When plotted the models might look like those of the upper plot of Figure 11.54 with six lines intersecting the y-axis at different points, but all having the same slope,
- second call to lmer: the operand  $(\text{frequency}-1 | \text{processor})$  specifies there is variation in the value of the slope and the source of this variation is the processor variable. When plotted the models might look like the lines in the middle of Figure 11.54, where all lines intersect the y-axis at the same point but have different slopes.
- third call to lmer: the operand  $(\text{frequency} | \text{processor})$  specifies that variation in the processor variable causes both the intercept and the slope to vary. When plotted, the models might look like the lines in the lower plot in Figure 11.54, where the lines have different intersections and slopes.

The following is the `summary` output from a mixed-effects model where the processor is a random effect on both the intercept and slope: `code`

```
Linear mixed model fit by REML [ 'lmerMod' ]
Formula: meanpower ~ frequency + (frequency | processor)
Data: power_bench
```

```
REML criterion at convergence: 6300.3
```

```
Scaled residuals:
    Min     1Q   Median     3Q    Max 
-4.0533 -0.4866  0.1453  0.4994  6.6743

Random effects:
Groups      Name        Variance Std.Dev. Corr
processor (Intercept) 0.12904  0.3592    
                  frequency  0.07383  0.2717   -0.99
Residual           0.10941  0.3308    
Number of obs: 9981, groups: processor, 6
```

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.1740    0.1473 14.76
frequency    1.9156    0.1111 17.23
```

```
Correlation of Fixed Effects:
  (Intr) 
frequency -0.993
```

The estimated coefficients, listed under `Fixed effects:`, for `(Intercept)` and `frequency` are very similar to the combined data model fitted earlier.

Annoyingly the `summary` output does not include p-values. These can be obtained using the `Anova` function from `car` package.

The `Random effects:` information lists the variation introduced by `processor` (listed in the `Groups` column, on the variables listed in the `Name` column); `Residual` lists the residual left after taking all the specified random effects into account.

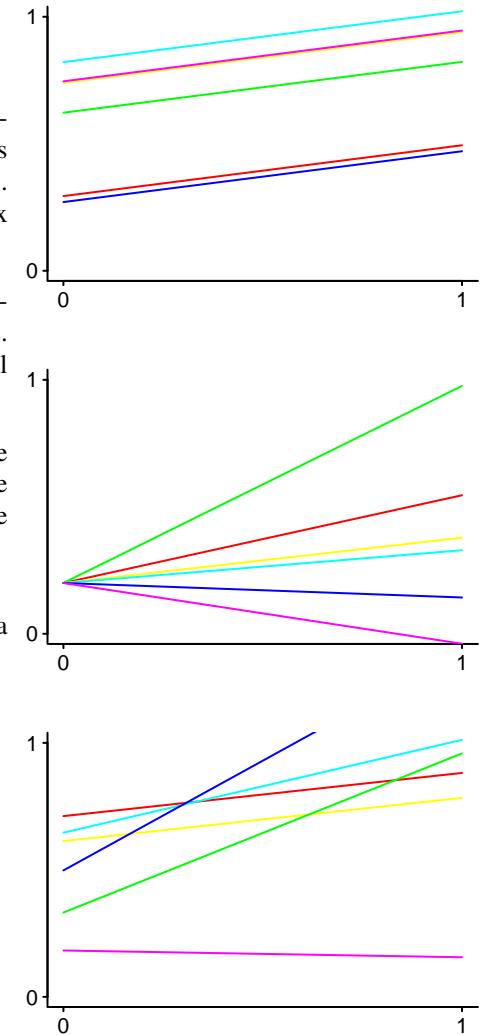


Figure 11.54: Example showing the three ways of structuring a mixed-effects model, i.e., different intersections/same slope (upper), same intersection/different slopes (middle) and different intersections/slopes (lower). `code`

Taking frequency as an example, there are two sources of uncertainty in its contribution to the response variable (as expressed in its coefficient), one from fixed effects and a random effect caused by processor variation.

Plotting the 95% confidence intervals for the intercept and slope of a mixed-effects model provides a visualization of the relative contribution of the sources of variation. Figure 11.55 was generated by the following code, using data from the six processors:

```
library("lattice")
library("lme4")
library("gridExtra")

proc_mod=lmer(meanpower ~ frequency +(frequency | processor),
              data=power_bench)
dp_orig=dotplot(ranef(proc_mod, condVar=TRUE), main=FALSE)

power_bench$shift_freq=power_bench$frequency-min(power_bench$frequency)
proc_mod=lmer(meanpower ~ shift_freq +(shift_freq | processor),
              data=power_bench)

dp_shift=dotplot(ranef(proc_mod, condVar=TRUE), main=FALSE)

# dotplot comes from the lattice package which uses grid layout
grid.arrange(dp_orig$processor, dp_shift$processor, nrow=2)
```

The upper plot is the model fitted using the original data; the intercept and slope (right plot) appear to be correlated. Looking at the straight lines fitted to each processor (Figure 11.53), they appear to share an origin starting at the lowest frequency measured; an intercept included as a random effect has a common origin assumed to start at zero (see Figure 11.54). Shifting frequency values down by the minimum measured value and refitting the model produces the confidence intervals in the lower plot. The correlation has disappeared; perhaps including the intercept as a random effect is not worthwhile.

Refitting a model without the intercept as a random effect, produces a model that only differs from previous models by a small amount (see [regression/hotpower-mix-plot](#)).

There is a limit on the number of random effects (i.e., number of unknowns) that can occur in a model. The total number of unknown random effects must be less than the number of observations, otherwise the equations do not have a unique solution. A continuous explanatory variable counts as a single unknown, while a variable holding nominal or ordinal values contributes one unknown for each of the possible discrete values (there is no slope associated with fitting a variable that is not treated as being continuous)...

Other operators ( $a \parallel b$ ) ( $a \mid b$  /  $c$ )...

The bootstrap can be used to calculate confidence intervals on a regression model. The `Boot` function in the `car` package will bootstrap the regression model passed as its first argument... [developers/jones\\_bin\\_prec.R](#)

## 11.7 Generalised Additive Models

The regression modeling techniques discussed so far have required the user to provide an equation expressing every detail of the relationship between explanatory variables and the response variable (they are said to be parametric models). If there is no reason to believe that any particular equation applies, then a *Generalised additive model* (GAM) provides an alternative approach. A GAM only requires a list of explanatory variables and a response variable to be specified (these models are said to be nonparametric models).

A GAM is built by finding the best fit for a sequence of polynomial equations (e.g., some form of spline) that smoothly captures the shape of the data. These smooth equations might be used to make predictions, or when the fitted model is plotted may suggest possible parametric equations. The details of the fitted equations are not a source of understanding, but they can make very good predictions.

The `gam` function in the `mgcv` package can be viewed as extending the functionality of `glm` to support a variety of nonparametric smoothing functions (the `gam` package is simpler, but

does not offer such a wide range of functionality). The following code shows formulas using a different smoothing function for each explanatory variable (first line below), a different smoothing function for some combinations of explanatory variables (second and third line), a combination of a smoothing function and parameterised form (fourth line) or an interaction between a smoothed and non-smoothed variable (fifth line; the `by` parameter rather than the `:` operator has to be used in this case):

```
mod=gam(y ~ s(x_1) + s(x_2) + s(x_3), data=foo_bar)
mod=gam(y ~ s(x_1) + s(x_2, x_3), data=foo_bar)
mod=gam(y ~ s(x_1) + s(x_2, x_3) + s(x_3, x_4) + s(x_4), data=foo_bar)
mod=gam(y ~ x_1 + s(x_2) + x_3, data=foo_bar, family="poisson")
mod=gam(y ~ x_1 + s(x_2, by=x_1) + x_3, data=foo_bar, family="poisson")
```

The smoothing function `s` supports a variety of options for controlling the fitting process. Two that are likely to be encountered are `k`, used to set an upper limit on the degrees of freedom that can be used in the fitted equation, and `bs` a string identifying the kind of smoother (e.g., "tp", the default, for a thin plate regression spline and "cr" for a cubic regression spline).

The value of `k` needs to be large enough to support the degrees of freedom needed by a polynomial capable of representing the underlying pattern in the data, but not to large as to require unacceptable computational resources. The `gam.check` function provides information about fitted models that can be used to help select a value for `k`.

The fitting procedure used by the `mgcv` version of `gam` tries to avoid overfitting by making every degree of freedom pay its way (using, for instance, *penalized regression splines*). Criteria used for measuring the *cost-effectiveness* of more complicated models include generalised cross-validation (GCV; the default) and AIC. The `select` argument provides support for *null space penalization*, see package documentation for details.

A study by Lee and Brooks<sup>826</sup> built a model to predict the performance and power consumed by applications running on processors having various hardware configurations (e.g., number of registers, size of cache and instruction latency).

The following additive model is based on the one proposed by Lee et al and explains over 95% of the variance in the data (see [regression/lee2006.R](#)). While this model is likely to be useful for prediction, it provides virtually no insight into the performance characteristics of the various hardware attributes.

```
l_mod=gam(sqrt(bips) ~ benchmark + fix_lat
           +s(depth, k=4) + s(gpr_phys, k=10)
           +s(br_resv, k=6) + s(dmem_lat, k=10) +
             s(fpu_lat, k=6)
           +s(l2cache_size, k=5) + s(icache_size, k=3) +
             s(dcache_size, k=3)
           +s(depth, gpr_phys, k=10)+s(depth, by=width, k=6)
           +s(gpr_phys, by=width, k=10)
           , data=lee)
```

An earlier example (Figure 8.31) showed two approaches to modeling the number of accesses to a function's local variables. Without knowing anything about what relationships might exist between explanatory and response variables, and being willing to use very high degree polynomials, it is possible to build and use `gam` to build a prediction model

In the calls to `gam` below, the first assumes there is an interaction between the two explanatory variables (allowing up to 75 degrees of freedom) and the second assumes the variables are independent (allowing up to 50 degrees of freedom for each of them). While the fitted model might make usable predictions (see [sourcecode/local-use/obs-fit.R](#)), the use of such high degree polynomials suggests that the underlying model has a non-polynomial form.

```
locg_mod=gam(norm_occur ~ s(object.access, total.access, k=75),
              data=common_loc, family=Gamma)

locp_mod=gam(norm_occur ~ s(object.access, k=50)+s(total.access, k=50),
              data=common_loc, family=Gamma)
```

## 11.8 Item response theory

The `ltm` package...

??

?

### 11.8.1 Agreement between raters

A measurement may be based on human judgement, e.g., making a product rating. Different people may make different judgements of the same characteristic/entity and a way of evaluating the agreement between the different judgements is needed.

Cohen's Kappa is a measure of inter-rater agreement between two raters, it varies from zero (no agreement) to one (perfect agreement). When comparing categorical variables, all differences are usually given equal weight (i.e., considered equally different). When comparing ordinal values, where the degree of difference between raters is easier to quantify, a weighting based on the square of the difference is often recommended...

Fleiss's Kappa is a measure of inter-rater agreement between three or more raters.

The `kappa2` function in the `irr` package supports Cohen's Kappa (weighting is supported by passing the argument `weight="squared"`). The `kappam.fleiss` function supports Fleiss's Kappa.

A study by Schach, Jin, Yu, Heller and Offutt<sup>1203</sup> categorised the kinds of maintenance activity performed on various systems. Table 11.4 lists the categories assigned by two raters to 215 maintenance categories involving the first 20 versions of Linux. The value of Cohen's Kappa for these two raters is 0.805 (see [group-compare/agreement.R](#)).

	Adaptive	Corrective	Perfective	Other	Total
<b>Adaptive</b>	2	0	0	0	2
<b>Corrective</b>	0	82	16	0	98
<b>Perfective</b>	0	5	99	2	106
<b>Other</b>	0	0	0	9	9
<b>Total</b>	2	87	115	11	215

Table 11.4: Maintenance categories assigned by two raters for the first 20 versions of the Linux kernel at the change-log level. Data from Schach et al.<sup>1203</sup>

Rater agreement as a means of measuring ability of rules to be accurately followed<sup>1402</sup>...

Raters with low levels of expertise, or few raters per item...?

## 11.9 Miscellaneous

Stuff that has no other obvious home...

The p-value, for the coefficients of a fitted model, is a test of the hypothesis that the coefficient is zero, i.e., there is no association. When the actual value of a coefficient is close to zero, the reported p-value may be spurious. One solution is to rotate the axes, which will have the effect of increasing the value of the coefficient and removing any artefact from the p-value calculation.

### 11.9.1 Advantages of using `lm`

Many books using R introduce readers to regression through the `lm` function; one reason for this is herd mentality, it's what everybody else does. While this book promotes `glm` as a one stop solution, the `lm` function does have some advantages over `glm`, including:

- requiring less cpu time to fit a model. For extremely large datasets, the performance difference may be worth considering,

- requiring less memory to fit a model. For extremely large datasets' memory requirements may be excessive for `glm`; possible solutions that continue to use `glm` are discussed below,
- the algorithm used by `lm` is always guaranteed to converge to a solution, singularities generated by correlation between explanatory variables excluded. There are edge cases where `glm` does not find a solution without being given some reasonable starting values...

The implementation of `lm` is based on the mathematics of *Ordinary Least Squares* (OLS) and the data has to satisfy more conditions for OLS to be applicable... constant variance...

The `lm` function requires that the error variance is constant (in practice close to constant). The `ncvTest` function, in the `car` package, checks that a fitted model meets this requirement; the `spreadLevelPlot` function provides some visualization. Also, see the `lmtest` function.

A user interface issue with models built using `glm` is that they do not come with an easy to understand goodness of fit number (`lm` has the R-squared value ... no r-squared supplied by `glm` and ???).

## 11.9.2 Network data

It is possible to build regression models over network data. But it does require some data to show how...

?

## 11.9.3 Alternative residual metrics

The default error metric used by many regression techniques is based on squaring the difference between the actual and predicted value. This choice is driven by the usefulness of the mathematical properties of sum of squares. Other error metrics could be used, for instance the absolute difference between actual and predicted values.

The `r1m` function in the `MASS` package supports user specified functions for calculating the residual to be minimised when fitting a model (the `psi`). Supported functions include... Huber...

## 11.9.4 Quantized regression

While removing 1% of outliers might make a significant difference to model accuracy, a quantized regression model might be more informative...

Minimises the residuals of the median (rather than the mean)...?

## 11.9.5 Prediction vs. interpretation

A different set of trade-offs...

interested in minimising local residual error when predicting, rather than global residuals when interpreting...

variance bias tradeoff...

## 11.9.6 Solving systems of equations

systems of unrelated equations, the `systemfit` package...

under-overspecified linear equations or inequality conditions `limSolve` package...

### 11.9.7 Very large datasets

The `biglm` package allows some kinds of regression models to be built using data that is too large to fit in memory all at once; the models are built using an incremental algorithm that only requires a subset of the data to be held in memory at any time. A variety of options are available for creating chunks of data to feed into the model building process, including incremental reading from files and databases.

The `biganalytics` package extends the `bigmemory` package by providing interfaces to various analytic packages, such as `biglm` (see [benchmark/bounds\\_chk.R](#))...

The `glm` function stores a lot of information in the object it returns, much of which is often not subsequently used...

### 11.9.8 Communicating model details

Information about fitted models needs to be communicated to the intended audience. The output produced by `summary` may appear cluttered in some contexts and it is not the easiest to interpret for casual users...

The `texreg` package contains function for mapping statistical model values to LaTeX and HTML tables...

## 11.10 Time series

Time series analysis deals with measurements that are sequentially correlated. An example of a measurement that is correlated with the previous measurement is current room temperature, which is likely to be similar to the temperature 10 minutes ago and the temperature in 10 minutes from now.

Techniques developed to analyse time-series can be used to analyse measurements of any quantity where a correlation exists between successive measurements.

A time series contains one or more of the following three components:

- underlying trend: which changes slowly,
- regular recurring pattern of changes (known as *seasonality*): for instance, expected daytime temperature throughout the year,
- random, irregular or fluctuating component.

The `stl` function (Seasonal Trend using Lowess) provides an easy way of splitting a time series (the argument must be an object of type `ts`, with a user specified frequency; the `stl` function does not automatically detect the recurrence period) into these three components (`plot` displays the individual components).<sup>xxvii</sup>

Figure 11.56, from a study by Eyolfson, Tan and Lam,<sup>402</sup> shows the three time-series components of the hourly rate of commits to the Linux kernel source tree, over the days of a week (the commits during the same hour of the same day were summed). The `stl` function assumes a fixed, recurring, pattern of seasonal behavior, a slowly changing trend and everything else is classified as random noise.

---

<sup>xxvii</sup> The `decompose` function, part of the base system, implements the same functionality in a less sophisticated way.

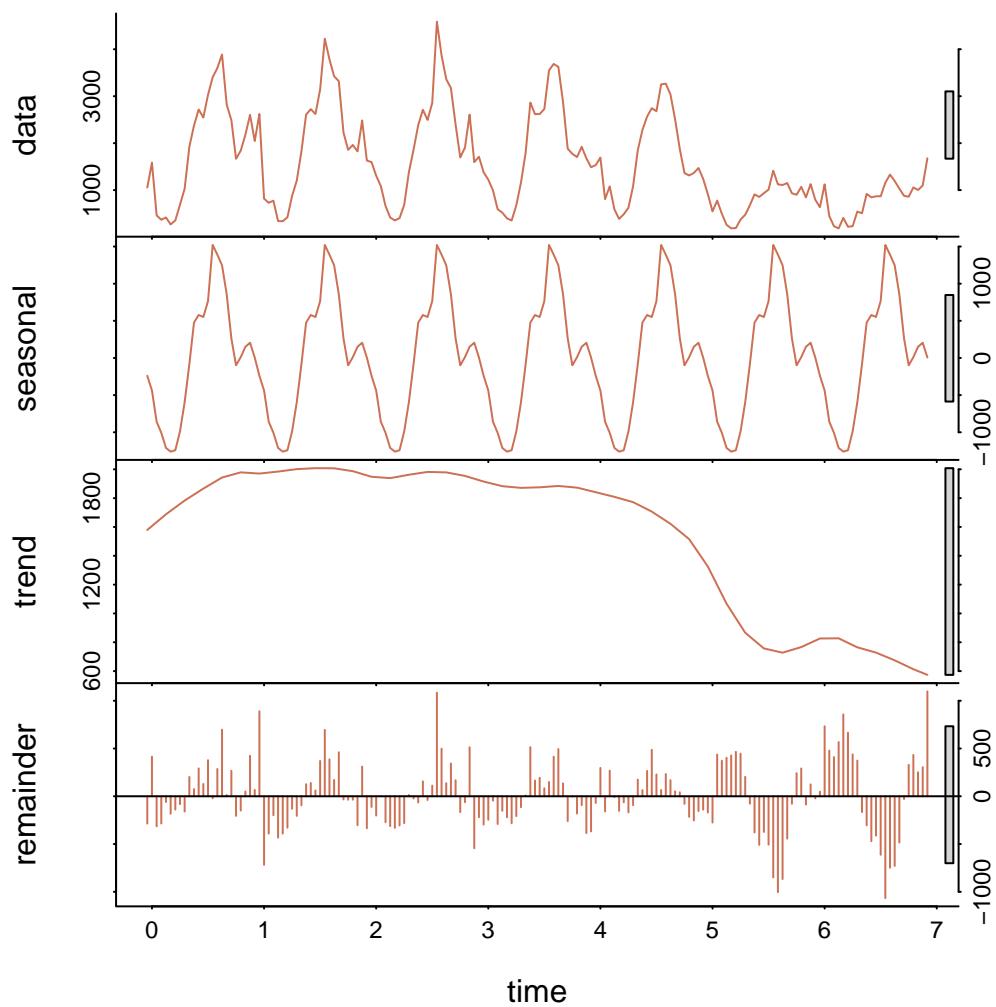


Figure 11.56: The three components of the hourly rate of commits, during a week, to the Linux kernel source tree; components extracted from the time series by `stl`. Data from Eyolfson et al.<sup>402</sup> [code](#)

```
# A seasonal frequency has to be specified
hr_ts=ts(linux_hr, start=c(0, 0), frequency=24)
plot(stl(hr_ts, s.window="periodic"))
```

Possible outputs from a time-series analysis include:

- a model of how the value of a quantity at time  $t$  depends on its values at an earlier time (often  $t - 1$ ),
- a regression model that take account of correlation between measured values,
- power spectrum showing the dominant frequencies present in the data,
- hierarchical clustering of multiple time series.

Structure is often added to the continuous, linear, nature of time by imposing repeating fixed length intervals, such as hours of the day and days of the week. Many time series analysis techniques require measurements to occur fixed length intervals; analysis of measurements at irregular intervals is not discussed here (but if we had some data...).

Some library functions use a time series datatype for representing time related measurements. The `ts` function, part of the base system, converts a vector to class `ts` (many time series functions will automatically convert vectors to this class).

The `xypoint` function, in the `lattice` package, can be used to create a time series strip chart, see Figure 8.18.

### 11.10.1 Cleaning time series data

Many time series techniques implicitly assume that measurement data occurs at regular intervals. A measurement process may only record events when they occur and if no event

occurred in within an interval there may be no data-point for that interval. Part of the cleaning process involves ensuring that every interval contains a value (which may be zero or inferred from surrounding values).

A study by Buettner<sup>198</sup> gathered project staffing information for various commercial development software projects. On large commercial projects the amount of work done at weekends is likely to be zero (except for the weeks prior to major deliveries) and the autocorrelation of project activity is likely to show a recurring pattern involving two consecutive days separated by seven days, i.e., weekends and weekdays.

Figure 11.57 shows the autocorrelation of the number of defects found on a given day for one development project. The seven day recurring pattern involves three consecutive days, are the developers only working a four-day week? It turns out <sup>xxviii</sup> that contractors on some projects work a two-week cycle, with extra hours worked one week and then not working the Friday of the following week.

The extent to which regular staffing level differences between Friday and other weekdays has to be taken into account will depend on the kind of analysis performed (weekends can be handled by excluding them from the analysis, focusing on where most effort occurs, i.e., week days).

Measurements made on public holidays, such as the New Year, are very likely to differ from normal work days. Removing public holidays from the data will scramble the association with day of the week. The extent to which day of the week is a more important factor in the analysis than public holidays has to be considered.

### 11.10.2 Modeling time series

The expected mean of a time series can be modeled using one or both of the following two approaches (samples containing values where the variance is serially correlated are discussed later):

- the *Autoregressive model* (AR) is based on the idea that the value at time  $t$  can be modeled as a weighted combination of values measured at earlier time steps, plus some amount of added noise ( $w_t$ ), for instance:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t$$

is an autoregressive model of order 2, abbreviated AR(2), because it uses values from two time steps back (with weights  $\phi_1$  and  $\phi_2$ ).

The `ar` function fits data to an autoregressive model.

- the *Moving Average model* (MA) is based on the idea that the value at time  $t$  can be modeled by as the sum of noise ( $w_t$ ) and a weighted combination of the noise from earlier time steps, for instance:

$$x_t = w_t + \theta_1 w_{t-1}$$

is a moving average model of order 1, abbreviated MA(1), because it uses noise (with weight  $\theta_1$ ) from one step back.

The `arima` function with the first two values of the `order` argument set to zero can be used to fit data to a moving average model.

In both of these models there is a correlation between the value of the response variable  $x_t$  and values of this variable at earlier time steps; for the AR(1) model  $x_t = \phi x_{t-1}$  the correlation between values separated by  $k$  time intervals has decreased by  $\phi^k$ . The autocorrelation function returns the correlation of a time series with itself at successive intervals (i.e., the correlation of the measurement at time  $t$  with the measurement at time  $t + n$ ;  $n=1:25$  is the default sequence of intervals); the `acf` returns and plots this function, see Figure 11.58.

The lag 0 autocorrelation is always one and the two dotted blue lines are 5% p-value bounds. Each lag is a hypothesis test and with 25 hypothesis tests (the default lag used) at least one value is expected to exceed a 5% p-value (with probability  $1 - 0.95^{25} \rightarrow 0.72$ ), also successive measurements are correlated and so neighbouring lag points are likely to show similar significance levels.

<sup>xxviii</sup> Email discussion with Buettner.

The partial autocorrelation function (implemented in the `pacf` function) calculates and plots the correlation at lag  $k$  after removing the effect of any correlation generated by terms at shorter lags. The partial autocorrelation at lag  $k$  is the  $k^{\text{th}}$  coefficient of a fitted AR( $k$ ) model; the `pacf` returns and plots this function, see Figure 11.59.

The previous two plots illustrate how short range correlations in an AR model have a long range impact on the values returned by `acf`, but an MA model does not have a long range impact, while the opposite behavior is seen in the values returned by `pacf`. An ARMA model always behaves in the most unhelpful way.

An ARMA model (*Autoregressive Moving Average*) is a combination of an AR and MA model, e.g., ARMA(2, 1) is the sum of an AR(2) and MA(1) model, such as the following:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + w_t + \theta_1 w_{t-1}$$

A time series is said to be *stationary* if the expected mean value does not change over successive measurements, i.e.,  $E[t_i] = E[t_{i+k}]$ , or in conceptual terms the probability of events driving a stationary process do not change over time. The mathematics behind both the AR and MA models assume a stationary time series.

ARIMA (*Autoregressive Integrated Moving Average*) handles non-stationary time series (implemented by the `arima` function) is a technique for handling certain kinds of non-stationary time series.

Many software engineering processes include components that change over time, e.g., more developers working on a project, more customers, larger systems, etc. Time dependent components having a significant impact on measured values create in a non-stationary time series. A variety of techniques are available for analyzing non-stationary time series (including converting them to a stationary form).

Time series analysis is not limited to data involving time, it can be used for any data that contains serial correlation between measurements.

A study by Hindle, Godfrey and Holt<sup>620</sup> measured the indentation of the first non-whitespace character on a line, for code written in a variety of languages. Figure 11.60 shows the autocorrelation of a list, ordered by indentation, of the total number of lines having a given indentation.

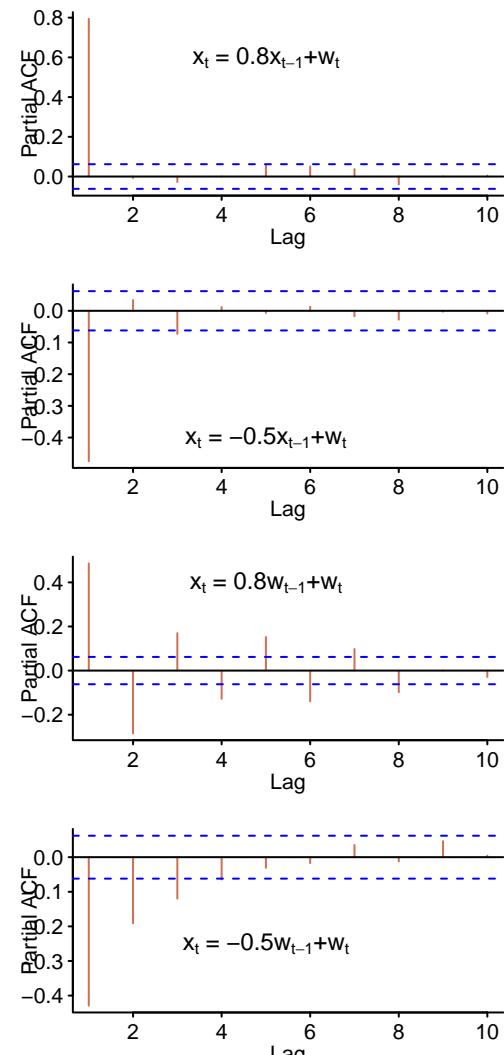


Figure 11.59: Partial autocorrelation of same two AR models (upper plots) and two MA models (lower plots) shown in Figure 11.58. [code](#)

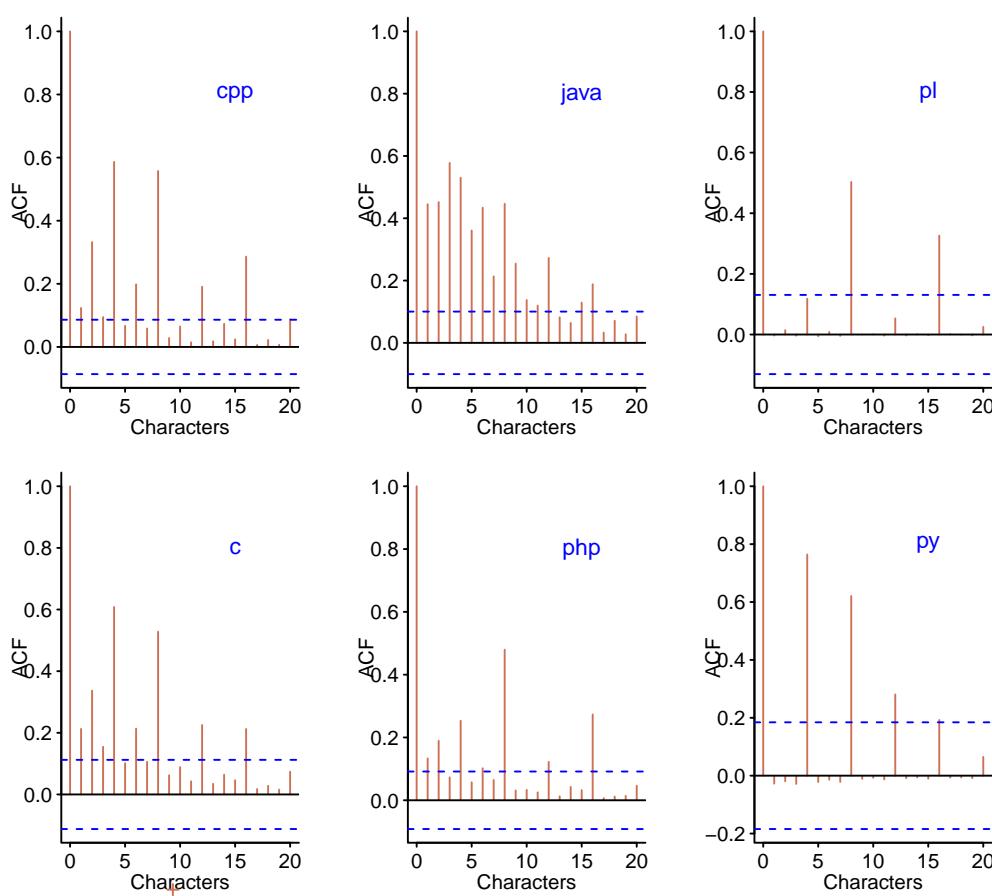


Figure 11.60: Autocorrelation of indentation of source code written in various languages. Data from Hindle et al.<sup>1160</sup>

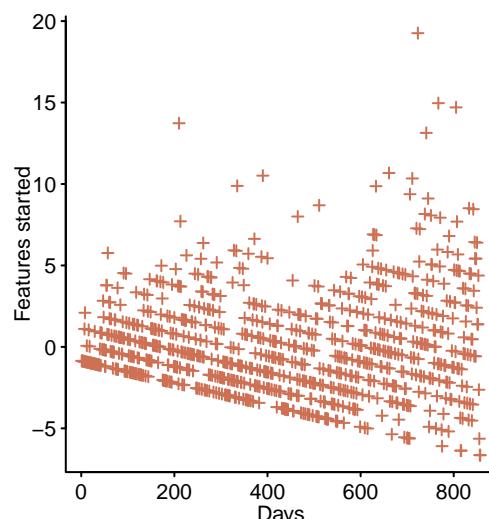
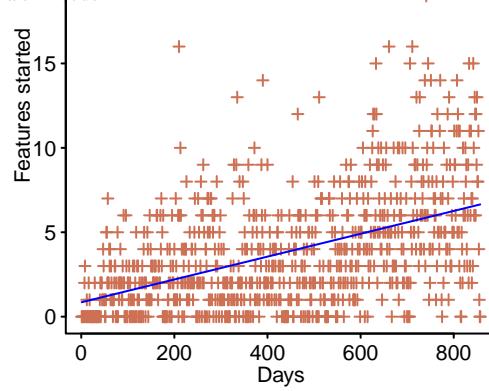


Figure 11.61: Number of features started for each day and fitted regression trend line (left) and number of features after subtracting the trend (right), over the entire period of the 7digital data. Data kindly supplied by 7Digital.<sup>1162</sup>

### 11.10.2.1 Building an ARMA model

ARMA modeling takes as input a time series and a non-stationary time series needs to be converted to a stationary form before model building can begin. Common reasons for a time series not being stationary and possible transforms to a stationary series include:

- a non-zero trend: for instance, the following equation contains an increasing time dependent trend:

$$x_t = \alpha + \beta t + w_t$$

Differencing can be used to remove trends, but care needs to be taken because this can introduce signals that are not in the original data. For instance, differencing the above equation gives:

$$\Delta x_t = x_t - x_{t-1} = b + w_t - w_{t-1}$$

an MA(1) process which the original series does not contain.

Subtracting the trend  $\alpha + \beta t$  leaves just  $w_t$  (see Figure 11.61),

- seasonality: this is a cyclic trend, e.g., changes that recur every year. Implementations of ARMA often support for including a seasonal component in the model, e.g., the seasonal to the arima function,
- non-constant variance (known as *volatility* in the analysis of financial time series).

If the growth in variance, over time, approximately follows the growth of the mean (perhaps there is a relatively consistent percentage change at each time step, e.g.,  $y_t = (1 + x_t)y_{t-1}$ ), then a log transform produces a time series with approximately constant variance (e.g.,  $\Delta(\log y_t) \approx x_t$ ).

A log transform can only be applied when values are greater than zero; in the case of the 7digital data (which has an increasing variance, as more developers are employed and time to implement features decreases) there are many zeroes and adding a tiny amount to every value allows a log transform to be made.

The plots produced by acf and pacf provide useful information about the likely structure and order of an ARMA model.

- if the plot produced by `acf` shows a decreasing trend, while the `pacf` shows a sharp cut-off (see Figure 11.58), an AR model is a good place to start,
- if the plot produced by `acf` shows a sharp cut-off, while the `pacf` shows a decreasing trend (see Figure 11.59), an MA model is a good place to start,
- if both plots show a decreasing trend, then some combination of AR and MA model is likely to be needed.

```
# Add 1e-5 to handle zero values
acf(diff(log(weekdays+1e-5)), xlab="Lag (working days)")
pacf(diff(log(weekdays+1e-5)), xlab="Lag (working days)")
```

The `arima` function can be called with various settings of the `order` parameter, and the quality of fits compared using AIC (arima only returns the series mean when difference value of zero is passed to `order`; the Arima function in the `forecast` package is not limited in this way). The `auto.arima` function in the `forecast` package can be used to find ARIMA model values that minimise AIC.

```
arima(diff(log(weekdays+1e-8)), order=c(5, 0, 1))
auto.arima(weekdays, max.order=7)
arima(diff(log(weekdays+1e-8)), order=c(1, 0, 2))
```

The two models that best fit for the feature start data are ARMA(5, 1) and ARMA(1, 2).

The output from the last call to `arima` above is: [code](#)

Call:  
`arima(x = diff(log(weekdays + 1e-08)), order = c(1, 0, 2))`

Coefficients:

	ar1	ma1	ma2	intercept
0.8577	-1.6663	0.6663	0.0102	
s.e.	0.0402	0.0579	0.0578	0.0022

`sigma^2` estimated as 46.75: log likelihood = -2862.8, aic = 5735.59

The Coefficients: table lists the model coefficients and their standard error. The `intercept` column is actually the time series mean (which for a stationary series is zero). The equation for one of the models is:

$$x_t - 0.0102 = 0.8577(x_{t-1} - 0.0102) + w_t - 1.6663w_{t-1} + 0.6663w_{t-2}$$

which simplifies to:

$$x_t = 0.0102 \cdot (1 - 0.8577) + 0.8577x_{t-1} + w_t - 1.6663w_{t-1} + 0.6663w_{t-2}$$

with the constant increment per time step evaluating to 0.00145146.

Which of these two possible models provides the best explanation of the data? Features take different amounts of time to implement and work can only start on a new feature when enough people have been freed through completion of work on other features. The coefficients of the AR component of the ARMA(5, 1) model can be interpreted as a probability that people working on a feature started a given number of days earlier will become available to start work on a new feature (see Table 11.5).

	AR	Duration
ar1	0.19	0.32
ar2	0.11	0.16
ar3	0.09	0.11
ar4	0.07	0.07
ar5	0.10	0.05

Table 11.5: AR coefficients of ARMA(5, 1) model and percentage of features taking a given number of days to implement. Data kindly supplied by 7Digital.[1162 code](#)

This may be a just-so story, but stories are useful tools and your author cannot think of one for the ARMA(1, 2) model.

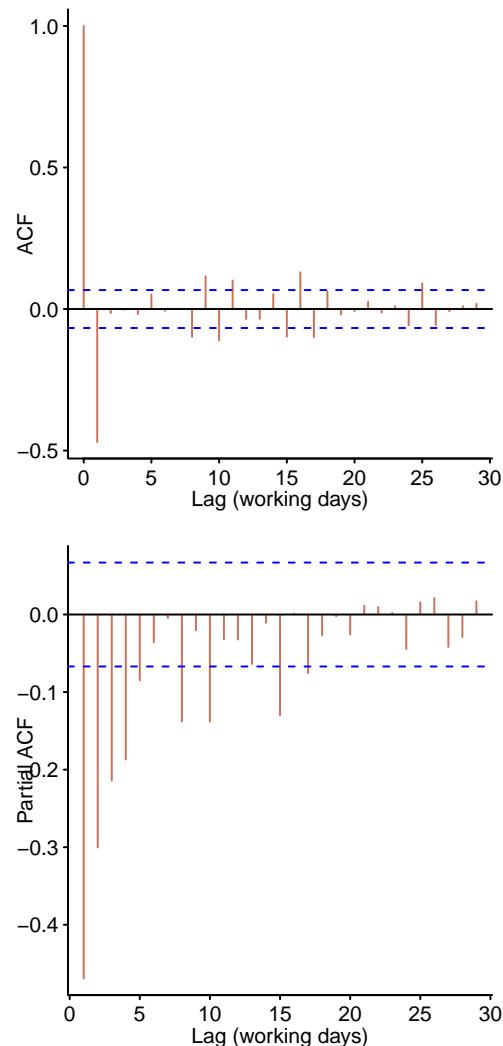


Figure 11.62: Autocorrelation (left) and partial autocorrelation (right) of the number of features started on a given day (after differencing the log transformed data), over the entire period of the 7digital data. Data kindly supplied by 7Digital.[1162 code](#)

**Handling seasonal trends** A seasonal ARIMA model can include AR, difference and MA components at an offset equal to the number of measurement intervals in the season.

By default, the `auto.arima` function, in the `forecast` package, will return seasonal components (if any are found). The `seasonal` option can be used to specify seasonal components to the `arima` function.

The following code estimates a seasonal ARIMA model for hourly commits to the Linux kernel source tree (see Figure 11.56):

```
library("forecast")

hr_ts=ts(linux_hr, start=c(0, 0), frequency=24)

auto.arima(hr_ts)
arima(linux_hr, order = c(2,1,1), seasonal = list(order = c(1,0,1), period=24))

The coefficients of the fitted models (below) differ because of differences in the algorithms used, but are within each other's standard error: code

Series: hr_ts
ARIMA(2,1,1)(1,0,0)[24]

Coefficients:
            ar1      ar2      ma1     sar1
            -0.9190  -0.3920  0.5022  0.6991
        s.e.    0.2059   0.0983  0.2186  0.0639

sigma^2 estimated as 141673:  log likelihood=-1233.55
AIC=2477.1   AICc=2477.48   BIC=2492.69

Call:
arima(x = linux_hr, order = c(2, 1, 1), seasonal = list(order = c(1, 0, 1),
period = 24))

Coefficients:
            ar1      ar2      ma1     sar1     sma1
            -0.8124  -0.2862  0.4483  0.8909  -0.4516
        s.e.    0.2995   0.1177  0.3058  0.0546   0.1404

sigma^2 estimated as 129070:  log likelihood = -1229.02,  aic = 2470.03

Call:
arima(x = linux_hr, order = c(2, 1, 0), seasonal = list(order = c(1, 0, 1),
period = 24))

Coefficients:
            ar1      ar2      sar1     sma1
            -0.3632  -0.1174  0.8980  -0.4727
        s.e.    0.1007   0.0964  0.0519   0.1391

sigma^2 estimated as 129942:  log likelihood = -1229.71,  aic = 2469.42
```

The `sar1` is the seasonal AR coefficient and `sma1` the seasonal MA coefficient.

The output from `auto.arima` is a suggested model. In this case the AR and SAR coefficients are pulling in opposite directions and the standard error for the MA coefficient is very high. Removing the MA component produces a model (second call to `arima` above) where the coefficients are not almost cancelling each other out; the model is (24 is the seasonal period):

$$x_t = -0.4x_{t-1} - 0.1x_{t-2} + 0.9x_{t-24} - 0.5w_{t-24}$$

What happened 24 hours ago is a better predictor than what happened in the previous hour or two hours.

**Predictions made using a fitted ARMA model** A fitted ARIMA model can be used to predict the values likely to occur after a measurement at time  $t$ . However, the relatively large noise component present in some ARMA models means that the confidence bounds of the predicted values quickly become very wide.

Figure 11.63 shows how the uncertainty in the ARIMA model built from the 7digital data swamps the predicted values.

Figure 11.63: Predicted daily difference in the number of new feature starts (red) and 95% confidence intervals (blue). Data kindly supplied by 7Digital.<sup>1162</sup> [code](#)

### 11.10.3 Non-constant variance

The variance of a time-series is primarily of interest when generating data for simulating processes. A time series that experiences rapid changes in variance is said to be *volatile*. Correlated variance is common during periods of volatility and techniques are available to build an autoregressive model for the variance, i.e. an *autoregressive conditional heteroskedastic* (ARCH) or a *generalised ARCH* (GARCH) model (a time series is *heteroskedastic* if the change in variance is regular and *conditionally heteroskedastic* if the change is irregular).

A time series experiencing changes in variance is not stationary.

An increase in frequency of commits leading up to a major new release is an example of behavior that can cause a change of variance in a time series.

The autocorrelation of a time series may show no correlation, but if its variance changes the square of the zero adjusted values will have a pattern of decreasing correlation in its acf, see lower plot of Figure 11.64.

```
acf(t_series)
acf((t_series-mean(t_series))^2) # Check for changing variance
```

The garch function in the tseries package fits a GARCH model to data.<sup>501</sup>?

### 11.10.4 Long-memory processes

Correlation at high lags... just need the data... Fractal nature of LAN traffic... (is not SE)

Fractionally differences ARIMA processes (FARIMA)...

The fracdiff package...

### 11.10.5 Smoothing and filtering

Smoothing a time series can make it easier to visually identify larger scale patterns and also provides a simple approach to predicting the immediate future values.

Even when data does not contain a systematic trend or seasonal effects (perhaps because they have been removed), it may still be possible to make a good estimate of immediate future values based on immediate past values.

Smoothing using the *exponentially weighted moving average* (EWMA; also known as *exponential moving average*, EMA) uses the following formula:

$$EMA_t = \phi x_t + (1 - \phi)EMA_{t-1}$$

where:  $x_t$  is the measured value of  $x$  at time  $t$  and  $\phi$  determines the amount of smoothing. The *exponential moving standard deviation* (EMS) is given by:

$$EMS_t = \sqrt{\phi EMS_{t-1}^2 + (1 - \phi)(x_t - EMA_t)^2}$$

EMA and EMS can be used to detect when a real-time data stream trends outside of pre-specified bounds.

Holt-Winters smoothing is a generalization of exponential smoothing that uses three parameters: estimated level, slope and seasonality; the HoltWinters function can be used to both estimate and apply these parameters.

A call to plot will display the three components of a time series based on the value returned by the HoltWinters function... Seasonal component can vary...

The filter function can be used to apply a linear filter to a time series.

TODO some examples...

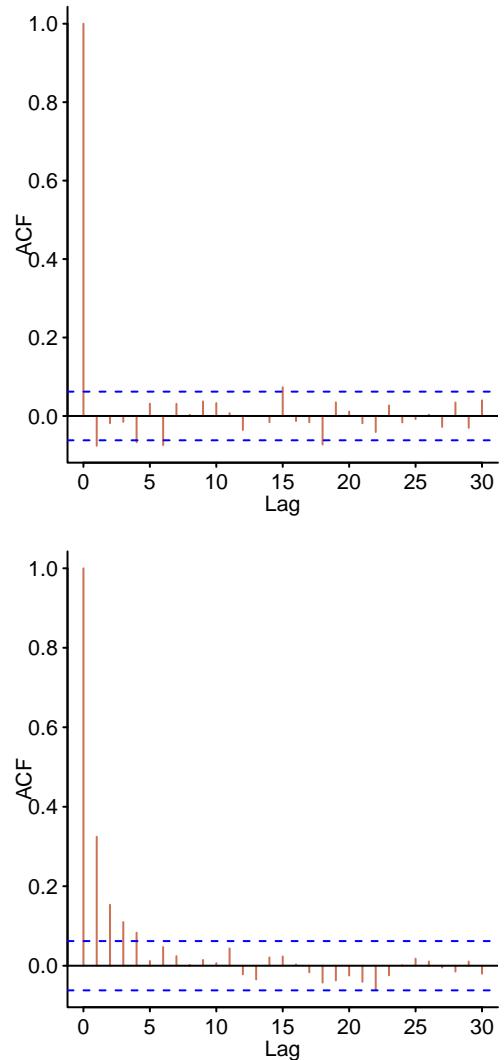


Figure 11.64: Time series whose values are uncorrelated (upper), but whose squared values are correlated (lower); see code for generation process. [code](#)

## 11.10.6 Missing data

Handling missing data in time series can be even more complicated and difficult than in regression modeling.

A study by Buettner<sup>198</sup> gathered data on project staffing and was not always able to obtain staffing information. Figure 11.65 shows a loess fit to the available data along with the standard error...

Some R functions support the use of splines for interpolation. Splines originated as a method for connecting a sequence of points by a smooth curve, not as a method of fitting a curve that minimises some error metric. Apart from their familiarity there is no reason to prefer the use of splines over other techniques (implementation issues also exist with the bs and ns functions in the splines package when building a model with the predict.glm function<sup>1398</sup> when making predictions using new data points)...

## 11.10.7 Spectral analysis

It is possible to transform a series of measurements in the time domain into the frequency domain. A stationary time series does not contain components at specific frequencies, but it can be described in terms of an average frequency composition.

The spectrum function (default calls spec.pgram function)...

The spec.arma function takes a specification of an ARMA model and returns its power spectrum, i.e., behaves like a call to spectrum when passed a time series that follows this model.

## 11.10.8 Relationships between time series

Some of the relationships that can exist between two or more time series include:

**Cross-correlation:** The correlation, at various lags, between two stationary time series. Figure 11.66 shows the cross-correlation between the number of source lines added/deleted, per week, to the glibc library. In calls to the ccf function, the first argument is the one which is shifted. In the following call:

```
ccf(lines_added, lines_deleted, col=point_col, xlab="Weeks")
```

the plot produces shows correlation spikes well above the confidence bounds occurring between the sequence pairs  $\text{lines\_added}_{t+2}/\text{lines\_deleted}_t$  and  $\text{lines\_added}_{t+8}/\text{lines\_deleted}_t$  (changes involving  $\text{lines\_deleted}$  is correlating with changes to  $\text{lines\_added}$  two and 10 weeks later; a positive lag means the first argument follows the second, a negative lag that it leads the second); there are small spikes at:  $\text{lines\_added}_{t-8}/\text{lines\_deleted}_t$  and  $\text{lines\_added}_{t-13}/\text{lines\_deleted}_t$ . Your author has no explanation for this correlation behavior.

**Alignment:** A time series is a sequence of values, with each value being larger, smaller or equal to the value immediately before it. If two time series are generated by the same, or similar, process they may contain subsequences of values that share the same pattern of up, down and don't change. A non-time series application of this kind of subsequence matching is locating word sequences common to two documents.

Dynamic time warping (DTW) is a class of algorithms that compares two series of values by stretching or compressing one of them (treated as the reference series) so it resembles the other (treated as a query series). The dtw package contains functions to perform and support DTW alignment of two series.

A study by Herraiz<sup>1333</sup> investigated the evolution of various long-lived software systems and measured the growth of NetBSD and FreeBSD (in lines of code). These two operating systems started from the same base, continue to share developers (see Figure 9.21) and code continues to be ported between them. Figure 11.67 shows the alignment, found by a call to dtw, between the weekly measurements of the lines of code in each OS.

```
library("dtw")

bsd_align=dtw(freebsd_weeks, netbsd_weeks, keep=TRUE,
              step=asymmetric, open.end=TRUE, open.begin=TRUE)
plot(bsd_align, type="twoway", offset=1, col=pal_col, xlab="Weeks")
```

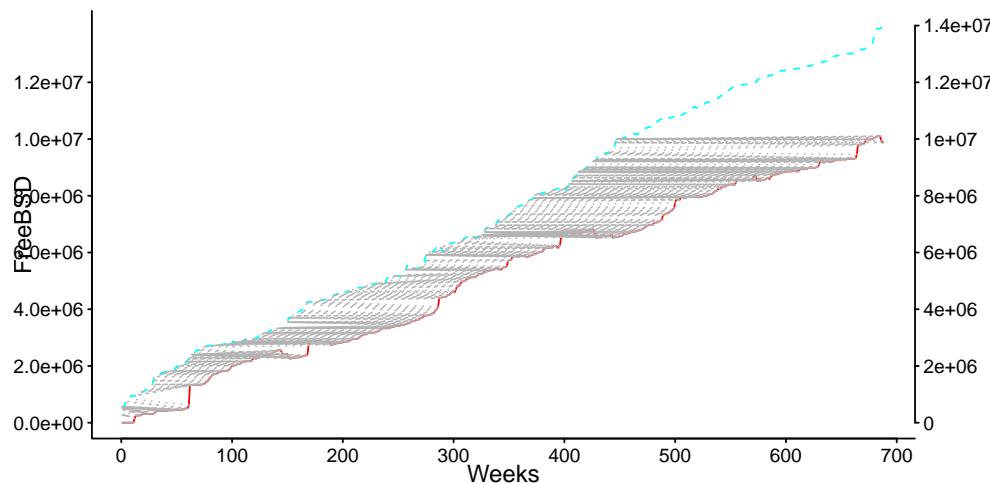


Figure 11.67: Visualization of alignment between weekly time series of lines code in NetBSD (blue) and FreeBSD (red). Data from Herranz<sup>1333</sup> [code](#)

**Clustering:** Many techniques for comparing two time series have been invented (at the time of writing the `diss` function, in the `Tsclust` package, supports 22 distance metrics). The pair-wise similarity of multiple time series can be used to cluster them by. The `Tsclust` package contains functionality for the clustering of time series.

A study by Powell<sup>1111</sup> investigated task effort allocation in a development project at Rolls-Royce. Figure 11.68 shows effort (in person hours) spent on eight major tasks (lower, from the bottom up: s/w requirements, top-level design, coding, low level test, requirement test, system acceptance test, management and holiday/non-project) and a hierarchical clustering of each task by its effort time series, with pair-wise distance between time series calculated using correlation (upper) and Euclidean (middle) metrics.

```
library("Tsclust")

eff_dist=diss(t(all_effort), METHOD="COR")
plot(hclust(eff_dist), main="", sub="", xlab="", ylab="Correlation distance")
```

### 11.10.9 Regression models

The mathematics underpinning many regression modeling techniques requires each measurement to be independent of the other measurements in a sample. Time series data is often serially correlated.

Some regression modeling functions can adjust for the presence of serial correlation (information about the correlation is passed in an optional argument). The `gls` function, in the `nlme` package, supports a `correlation` option; the `dynlm` package supports the use of time series operators (e.g., `diff` and `lag`) in the specification of model formula; the `tscount` package supports the fitting of generalized linear models to time series of count data.

[time-series/agile-week-acf.R...](#)

The number of source lines in FreeBSD is growing over time, see Figure 11.2...

### 11.10.10 Misc

Outlier detection...

Series containing irregular and regular processes?

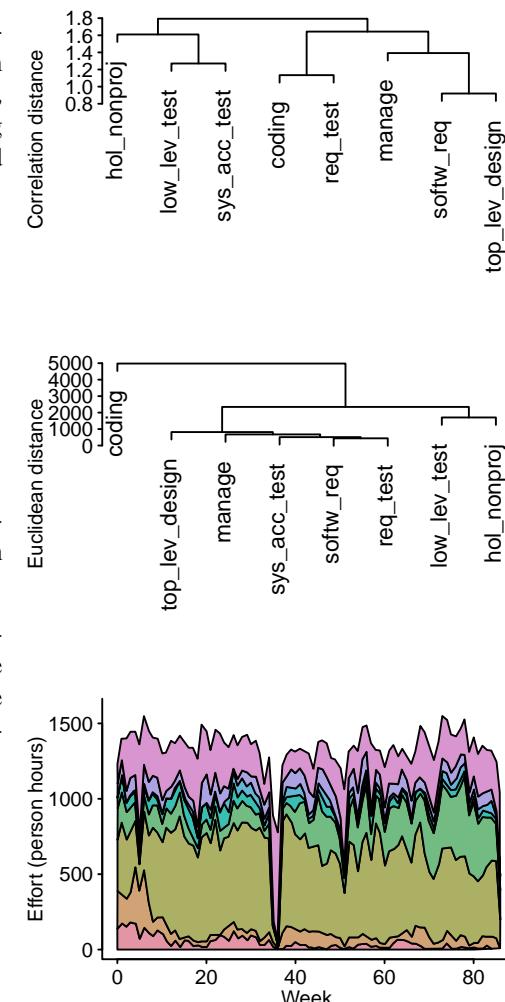


Figure 11.68: Effort distribution (person hours) over the eight main tasks of a development project at Rolls-Royce and a hierarchical clustering of each task effort time series based on pair-wise correlation and Euclidean distance metrics. Data extracted from Powell.<sup>1111</sup> [code](#)

The `cpm` package supports change-point analysis of time series having a variety of distributions, e.g., exponential, poisson and the `ecp` package supports change-point analysis of multivariate time series...

Granger causality...

## 11.11 Survival analysis

Survival analysis is the analysis of data where the response variable has the form of *time-to-event*. Historically this kind of model building has been used to compare the impact of different medical procedures, or drugs, on subject survival rate. In some cases the event of interest may not occur during the measurement period, the measurements in this case are said to be *censored*. A software example of censoring is measuring the time interval between a function definition being written and the first time it is modified; the measurement data is said to be *right censored* when one or more functions are not modified before the study ends.

By default, the analysis deals with one kind of event, which causes a transition to a terminal state, e.g., there is no coming back from the dead. Competing risk models deal with the situation where one of several risk events can cause the transition to the final state. Multistate models handle the situation where some transitions are to states that are not final, i.e., an appropriate event can cause a transition to another state.

Survival analysis makes greater use of the available information to produce estimates containing less error than other forms of regression modeling; a linear regression model comparing mean time-to-event between groups would have to ignore censored data, while a logistic regression model, using 0/1 to indicate whether a subject survived or not, would again have to ignore censored data.

Possible outputs from survival analysis include:

- survival function,  $S(t)$ , the probability of surviving a given amount of time, is used to estimate time-to-event for a group of subjects or compare time-to-event between subjects in two or more groups,
- hazard function,  $h(t)$ , the hazard rate, that is, the probability of an entity surviving to time  $t$  experiencing an event in the next time interval, e.g., having survived 69 years 11 months before reading this sentence the probability that you die in the next month (the interval used to denote an instant is small compared to the time spans involved). The survival and hazard functions can be derived from each other:

$$h(t) = \frac{f(t)}{S(t)}$$

where:  $f(t)$  is a probability density function, the probability of the event occurring at exactly  $t$  time units in the future, e.g., the probability of a baby born 70 years ago living long enough to read this sentence but not before,

- a regression model showing the impact of explanatory variables on time-to-event. This may be a non-parametric model, such as the Cox proportional hazard model, because parametric models can be very difficult to build.

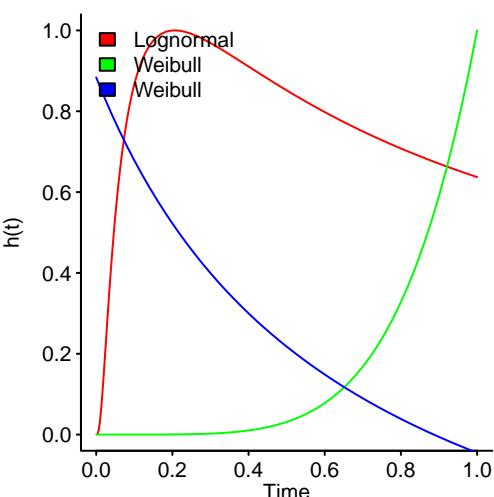


Figure 11.69: Two commonly used hazard functions; Weibull is monotonic (always increases, decreases or remains the same) and Lognormal which can increase and then decrease. [code](#)

Time-to-event is always positive and so has a skewed distribution (which means it cannot have a Normal distribution).

The `survival` package contains functions implementing the functionality needed to perform survival analysis.

### 11.11.1 Kinds of censoring

Ideally censoring is uninformative, i.e., the distribution of censoring times provides no information about the distribution of survival times. When a period of study is decided in advance, the censoring information is uninformative.

When censoring is not under the control of the experimenter, it is said to occur at random. For instance, a subject may decide to stop taking part in a study because they are not happy with their performance.

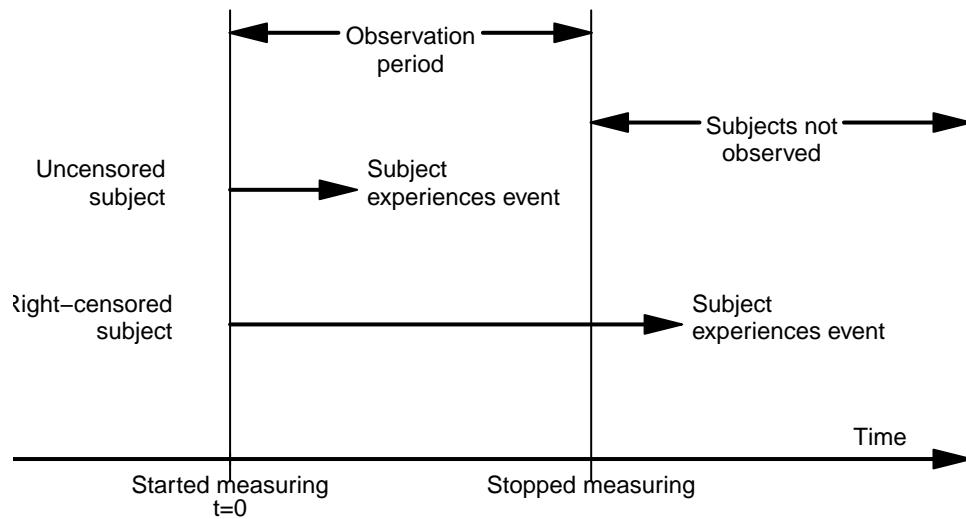


Figure 11.70: Observation period with events inside and outside the study period. [code](#)

The situations where censoring can occur include:

*Left truncation*: subject not observed before  $t_0$  experienced an event before that time and is not included in the study (the event may have been such that it rendered the subject unable to join the study, e.g., developer left the company),

*Left censored* (also *left truncated*): occurs when a subject included in the study is known to have had the event prior to time  $t$ , but with the exact time not being known.

*Right censored*: described at the start of the subsection,

*Interval censored*: when measurements are made at regular intervals, the exact time of an event is not known, only that it occurred between two measurement points,

*non-detect*: the measurement process may fail to detect an event because the strength of the event is below the detection threshold. This kind of censoring is not covered here, see Helsel.<sup>598</sup>

### 11.11.1.1 Input data format

The `Surv` function creates a survival object from data and the object it returns plays the role of the response variable in formula passed to model building functions. The required data format depends on the kind of censoring and presence of time dependencies. The following is an example of the basic information required:

```
id,start_time,end_time,failure_status,explanatory_v1,explanatory_v2
```

where: `id` is a unique identifier denoting each subject (only needed when information on the same subject occurs on multiple lines), `start_time/end_time` the starting time (or date) of measurement and the end time (either when the event occurred, the end of the study or the last recorded time of a subject who was not seen again) and `failure_status` one of two values specifying whether an event occurred or not; followed by an optional list of explanatory variables.

The time of interest is the difference between the start/end time and the data may contain just this value.

### 11.11.2 Survival curve

The *Kaplan-Meier* curve is a descriptive statistic of time-to-event measurements, that can include censored data. It shows the percentage of subjects who have not experienced an event up to a point in time and an optional confidence interval.

A study by Businge, Serebrenik and van den Brand<sup>206</sup> investigated the number of releases of Eclipse third-party plug-ins (ETP) between 2003 and 2010; the history of each ETP was traced from the year of its first release and any releases in subsequent years were noted.

The Eclipse framework includes a published list of officially recognised APIs and each release of the Eclipse SDK also includes support for APIs considered to be for internal API use, i.e., not applications. The status difference between official/internal APIs is that internal APIs can be changed without notice, while the official APIs are intended to have some degree of permanence (they may change on major releases but are not intended to change on minor releases; starting in 2004 all yearly releases were minor).

At some point there are no new releases of an ETP in a year and this cessation of new releases could be regarded as the *death* of development of the ETP (some ETP development died for one year only to be resurrected the following year; for simplicity the small number of such recurring events are ignored).

For this analysis ETP yearly release counts are divided into two groups, those that only made use of official APIs and those that made use of one or more internal APIs; Table 11.6 shows the number of ETPs using only the official API.

	2003	2004	2005	2006	2007	2008	2009	2010
2003	35	10	3	1	1	2	0	0
2004	0	33	4	4	2	2	0	0
2005	0	0	41	10	4	3	1	1
2006	0	0	0	61	7	1	0	2
2007	0	0	0	0	37	12	4	6
2008	0	0	0	0	0	38	7	2
2009	0	0	0	0	0	0	25	3
2010	0	0	0	0	0	0	0	16

Table 11.6: Total number of distinct ETPs released in a year; left column lists year of first release and releases in subsequent years. Data from Businge et al.<sup>206</sup>

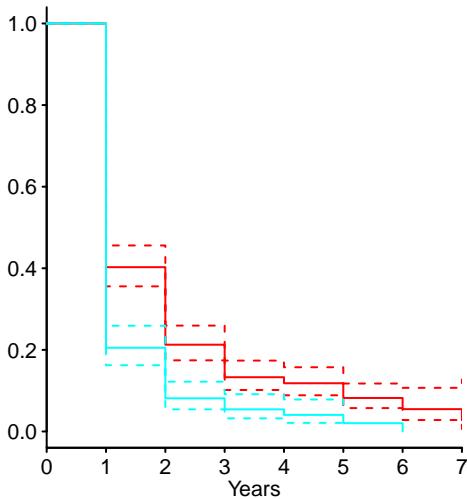


Figure 11.71: The Kaplan-Meier curve for survivability of new releases: (blue) ETPs using only official APIs, (blue) ETPs calling internal APIs (red); dotted lines are 95% confidence intervals. Data from Businge.<sup>205</sup> code

Figure 11.71 shows the Kaplan-Meier curve for ETPs using only official APIs (blue) and ETPs that use internal APIs (red); the dotted lines are 95% confidence intervals.

The plot was created as follows:

- calling the `Surv` function to create a survival object containing time and censored information on each subject (this is the first step in most survival analysis when using R),
- calling the `survfit` function with a formula containing the object returned from `Surv` as the response variable and the explanatory variable `API`,
- calling `plot` (or rather the overloaded version) with the model returned by `survfit`:

```
library("survival")
api_surv=Surv(all_API$year_end-all_API$year_start,
              event=(all_API$survived == 0), type="right")
api_mod=survfit(api_surv ~ all_API$API)
plot(api_mod, col=pal_col, conf.int=TRUE, xlim=c(0,7), xlab="Years")
```

The `summary` function can be used to obtain the values of the survival curve at each time measurement point.

**Comparing two survival curves** Are the two survival curves statistically different? The `survdiff` function can be used to answer this question. The p-value returned by the call below (bottom right) shows that the two survival curves are very unlikely to be the same: code

Call:

```
survdiff(formula = Surv(year_end - year_start, event = (survived ==
0), type = "right") ~ API, data = all_API)
```

N Observed Expected (O-E)^2/E (O-E)^2/V

API=0	381	334	372	3.83	29
API=1	289	260	222	6.41	29

Chisq= 29 on 1 degrees of freedom, p= 7.17e-08

By default, `survdiff` performs a *log-rank test*, which gives equal weight to all events. Passing the argument `rho=1` causes greater weight to be given to earlier events, while the argument `rho=-1` gives greater weight to later events. The hazard function is returned by `survfit` functions when it is passed the argument `type="fh"`.

Why, on average, do new releases of an ETP using internal APIs occur over a greater number of years? Is it because there are changes to the internal APIs that break the ETP, requiring the ETP to be updated to handle the change and a new version released, or is it because authors who use internal APIs are more committed to creating the best possible product and so continue to refine their ETP over more years?

Perhaps suspecting that changes to the SDK were a significant factor, Businge<sup>205</sup> investigated the source compatibility of ETPs with the Eclipse SDK across releases 1.0 to 3.7 (i.e., releases in every year from 2001 to 2011). Every ETP was built using each of these 11 SDK releases (yes, even SDKs created before an ETP was first released). To allow easy comparison with the ETP analysis above, the following analysis only considers SDK builds released after an ETP was first made available (see Table ?? for numeric values).

The Kaplan-Meier curve in Figure 11.72 shows the survival of ETPs' ability to build under the Eclipse SDK released in each successive year. ETPs using internal APIs (red) are much more likely to fail to build (precompiled plug-ins may still function if they don't call any changed internal API) when a new Eclipse SDK is released than ETPs using only the official APIs (blue).

Figure 11.72 suggests that developers using internal APIs in their ETP are more likely to be forced to release an update if they want their ETP to continue to function with later releases. However, this data does not address the possibility that developers who make use of internal APIs are more committed to creating the best possible product.

The median is preferred over mean as the measure of central tendency for survival data, because the mean underestimates the true value when samples contain censored data. The median is measured as the point where the Kaplan-Meier curve falls before 0.5 and printing the model returned by `survfit` gives this value along with its 95% confidence intervals.

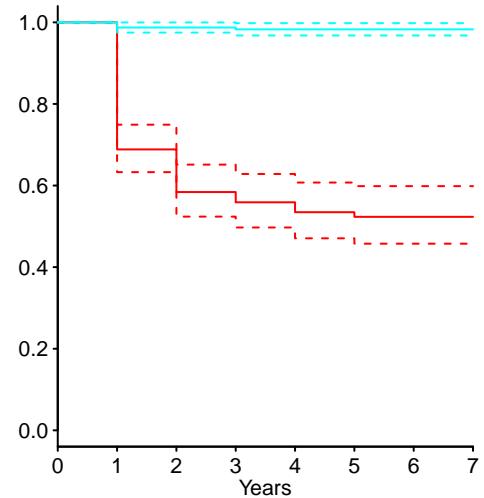


Figure 11.72: The Kaplan-Meier curve for survivability of ETPs ability to be built using SDK released in subsequent years: (blue) ETPs using only official APIs, (red) ETPs calling internal APIs; dotted lines are 95% confidence intervals, with plus signs, +, indicating censored data. Data from Businge.<sup>205</sup> [code](#)

### 11.11.3 Regression modeling

Survival data implicitly contains information that is not present in ordinary regression modeling; the probability of an event occurring at a given time, the hazard function. Estimating the appropriate hazard function for survival data requires knowing the coefficients of the explanatory variables in the regression model, and estimating the coefficients of the explanatory variables requires knowing the hazard function.

Model building functions will attempt to model the shape of the hazard function that is specified and if the chosen hazard function is incorrect, the returned model may be substantially incorrect. Also, parametric models have been found to be very sensitive to the explanatory variables provided as input to the model building process.

There is no single statistic available for definitively selecting the best model (i.e., hazard function and appropriate explanatory variables).

The Cox proportional-hazards model does not require the specification of a hazard function, breaking the circularity in selecting regression coefficients and removing some of the dangers associated with use of an incorrect hazard function (the Cox modeling approach is not guaranteed to always build a reasonably accurate model). If there is any doubt about the appropriate parametric distribution, the Cox model is a safe choice.

While the Cox proportional hazards model has many advantages, a potentially big disadvantage is that without specifying a hazard function it is not possible to make predictions outside of the interval covered by the measurements.

The `censReg` package supports fitting regression models to censored data...

### 11.11.3.1 Cox proportional-hazards model

A Cox proportional-hazards regression model provides reasonably good estimates for the coefficients of the explanatory variables and hazard ratios (not absolute values, but ratios) for a wide variety of data. The Cox model is popular because it is robust, it will closely approximate the correct parametric model. If the correct parametric model has a Weibull hazard function (whose shape parameter is unknown), the Cox model will give similar results to those obtained from this parametric model; if the parameters of the Weibull hazard function are known, a model built using them will outperform a Cox model.

The Cox likelihood (known as a partial likelihood) is based on the observed order of events rather than the interval between them (so it only considers subjects' experiencing an event).

The equation for the basic Cox model is (note that time is not yet included as an explanatory variable,  $x_{ki}$ ; the variables in this basic Cox model cannot be time dependent):

$$h_i(t) = h_0(t)e^{\beta_1x_{1i} + \dots + \beta_kx_{ki}}$$

where:  $h_i(t)$  is the hazard function for subject  $i$  at time  $t$ ,  $h_0(t)$  is a baseline hazard function and the contents of the exponent expression are explanatory variables and their regression coefficients ( $\beta_0$  is included as part of the baseline hazard).

This equation can be written as a log ratio of the hazard functions:

$$\log \frac{h_i(t)}{h_0(t)} = \beta_1x_{1i} + \dots + \beta_kx_{ki}$$

or as a hazard ratio for two subjects,  $i$  and  $j$ :

$$\frac{h_i(t)}{h_j(t)} = e^{\beta_1(x_{1i}-x_{1j}) + \dots + \beta_k(x_{ki}-x_{kj})}$$

Thus the Cox model assumes the effect of each explanatory variable is multiplicative.

The `coxph` function, in the `survival` package, builds Cox proportional-hazard models; the basic usage follows the pattern used by `glm`, with the object returned by `Surv` playing the role of the response variable. For example:

```
p_mod=coxph(Surv(patch_days, !is_censored) ~ log(cvss_score)+opensource,
            data=ISR_disc)
```

The `cox.zph` function can be used to check the assumption that the explanatory variables are not time dependent (at least during the measurement period).

If two or more events occur at the same time the associated data is said to be *tied*. The default value of the option `ties="efron"` option, can handle some tied data, but if many events occur at the same time (e.g., the ETP data in Table 11.6) calls to `coxph` might need to use `ties="exact"`.

The techniques for formula specification and refinement used with `glm` can also be applied to models created with `coxph`, e.g., starting with a complicated model and using `stepAIC` to simplify it.

A study by Arora, Krishnan, Telang and Yang<sup>55</sup> investigated the time it took vendors to release patches to fix vulnerabilities reported in their product; explanatory variables included information about the software vendor and whether the vendor was privately notified about the vulnerability or the vendor first found out about it through a public disclosure.

The following is the summary output from a model fitted by `coxph` to the data for public disclosure vulnerabilities: `code`

Call:

```
coxph(formula = Surv(patch_days, !is_censored) ~ log(cvss_score) +
      opensource + y2003 + smallvendor + small_loge + log(cvss_score):y2002 +
      y2002:smallvendor + y2003:smallvendor, data = ISR_np)
```

```
n= 945, number of events= 824
```

	coef	exp(coef)	se(coef)	z	Pr(> z )
log(cvss_score)	0.23283	1.26217	0.08570	2.717	0.00659 **
opensource	0.42235	1.52555	0.09167	4.607	4.08e-06 ***
y2003	0.83643	2.30811	0.10459	7.997	1.22e-15 ***

```

smallvendor      -0.40940  0.66405  0.17331 -2.362  0.01816 *
small_loge       0.02926  1.02969  0.01346  2.173  0.02975 *
log(cvss_score):y2002  0.23048  1.25920  0.04961  4.646 3.39e-06 ***
smallvendor:y2002  0.59685  1.81638  0.19540  3.054  0.00226 **
y2003:smallvendor  0.58999  1.80396  0.22502  2.622  0.00874 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
log(cvss_score)      1.262     0.7923   1.0670   1.4930
opensource            1.526     0.6555   1.2747   1.8258
y2003                 2.308     0.4333   1.8803   2.8332
smallvendor           0.664     1.5059   0.4728   0.9326
small_loge             1.030     0.9712   1.0029   1.0572
log(cvss_score):y2002  1.259     0.7942   1.1425   1.3878
smallvendor:y2002      1.816     0.5505   1.2384   2.6640
y2003:smallvendor      1.804     0.5543   1.1606   2.8039

Concordance= 0.647 (se = 0.012 )
Rsquare= 0.19 (max possible= 1 )
Likelihood ratio test= 199 on 8 df,  p=0
Wald test            = 184.9 on 8 df,  p=0
Score (logrank) test = 198.3 on 8 df,  p=0

```

The first half of the output is very similar to the `summary` output produced by a model fitted by `glm`. The table of numbers in the middle are 95% confidence intervals, which are printed by default.

The numbers in the bottom part of the table are the R-squared of the fit<sup>xxix</sup> (0.19 in this case, showing that only a small amount of the variance in the data is described by the model) and p-values for various tests of the null hypothesis that the coefficients are zero (abbreviated to a single letter, p).

The Cox model is a proportional-hazards model, the explanatory variable coefficients are proportions not absolute values. The coefficients specify the expected impact of the respective explanatory variable when the values of all the other variables are kept constant. On their own they cannot be used to predict response variable values, they can only be used to predict changes to known values.

Taking `log(cvss_score)` as an example, the value 1.26217 appears in its `exp(coef)` column. A  $\pm 1$  change in the value of `log(cvss_score)` is expected to change the response variable (time taken to produce a patch) by  $\pm(1.26217 - 1) \cdot 100 \rightarrow \pm 26.21$  percent (a value of less than one in the `exp(coef)` column reverses the sign of the percentage change, e.g., an increase in the value of the explanatory variable is predicted to decrease the value of response variable).

Model adequacy can be checked using Cox-Snell residuals and influential observations can be checked for using *score residuals*, which specify how each regression coefficient would change if a particular observation was removed (see [survival/vulnerabilities/patch-cph.R](#)).

**Frailty of subjects** The above analysis of time-to-patch has implicitly assumed that there is no difference between vendors in their ability to respond and fix reported vulnerabilities. Unobserved differences in subject performance means that there will be some variation in their hazard and *frailty* is the term used to denote these random changes in the hazard function. The effect of introducing the uncertainty implied by frailty, to a Cox model, is to add a random effect,  $v_j$ , the frailty of group  $j$  that  $x_i$  belongs to:

$$h_i(t) = h_0(t)v_j e^{\beta_1 x_{1i} + \dots + \beta_k x_{ki}}$$

The *frailty* function can be included in a formula to specify explanatory variables that identify particular groups of subjects sharing the same frailty. In the case of the vulnerability study, vendors are treated as the frailty group:

```
fp_mod=coxph(Surv(patch_days, !is_censored) ~ log(cvss_score)+opensource
               +frailty(vendor), data=ISR_disc)
```

The `summary` output includes the information: Variance of random effect=0.374 (see [survival/vulnerabilities/patch-frailty.R](#)).

<sup>xxix</sup> The value printed is the Cox & Snell pseudo R-squared, which can be less than one and the maximum possible value for the data appears in the `summary` output.

The  $v_j$  in the above equation is assumed to have a mean=1 and a variance that is calculated as part of the model building process (in this case it is 0.374). The main consequence of including frailty in a Cox model is to explicitly allocate some of the variance present in the data to a specific explanatory variable (the model coefficients of explanatory variables may also change).

The `frailtypack` package provides a wider range of frailty related options and functionality than is available in the `survival` package.

30 ...

### 11.11.3.2 Time varying explanatory variables

The behavior of explanatory variables may change over time; the options are either to exclude all affected subjects from the analysis or to use a technique that handles the time dependent behavior

The Arora et al study investigated the impact of public disclosure of vulnerabilities on the time it took vendors to release patches for their product. Possible event sequences were:

- vendor was privately notified about a vulnerability and some time later a simultaneous announcement of the vulnerability and a vendor patch was made (213 of 755 private notifications),
- vendor was privately notified about a vulnerability, but information about the vulnerability was made public before a patch was available for release (the vendor's patch being released some time later in 542 of 755 private notifications); this is a time dependent change of a significant attribute,
- the vendor learned about a vulnerability when information about it was made public, and sometime later released a patch (945 cases),

If privately notified and public disclosure fix rates are compared using a Kaplan-Meier curve, any privately notified vulnerabilities that become public before a patch is available have to be treated as censored (simply ignoring them biases fix rates towards a lower value; see Figure 11.73).

The first Cox model for the vulnerability data only used information for the case where the vendor found out about the vulnerability via public disclosure.

Building a regression model based on all the vulnerability data requires handling time dependent explanatory variables, which requires reformatting the data to make the time dependencies explicit. The time dependency, for this data, is a possible change of state from the vulnerability not being public to the information being public.

The original data looks something like the following:

```
notify,publish,patch,vendor,employee,os
2000-10-16,2000-11-18,2000-12-20,"abc",1000,unix
```

Publication of vulnerability information occurs before a patch is released and five columns are added, one to uniquely identify each vulnerability, the start/end dates the interval during which the information was private or disclosed, a flag each to specify private and disclosed, and whether an event (i.e., release of a patch) occurred in the interval. The first interval starts on the date the vendor was notified and ends on the date the vulnerability is made public, a second interval occurs for vulnerabilities that change state from private to disclosed before a patch is available and states on the date of disclosure and ends on the date a patch became available, as follows:

```
id,start,end,priv_di,notify,publish,patch,event,vendor,os
1,2000-10-16,2000-11-17,1,2000-10-16,2000-11-18,2000-12-20,0,"abc",unix
1,2000-11-18,2000-12-20,0,2000-10-16,2000-11-18,2000-12-20,1,"abc",unix
```

Treating `pr_di` as an explanatory variable (1 for private disclosure to vendor and zero for public disclosure) enables the impact of disclosure on patch time to be included in a model.

When all the measurement data has to be split on the same date, the `survSplit` function can be used to create the necessary rows, otherwise (as in this case) specific data mangling code has to be written.

The call to `coxph`, or `survreg`, has to include the term `cluster(id)`, which ties together (by vulnerability id in this case) the rows associated with the same subject. The call to `coxph` looks something like the following:

```
td_mod=coxph(Surv(patch_days, !is_censored) ~ priv_di*cvss_score
              +cluster(id), data=ISR_split)
```

It is not possible to use `cluster` and `frailty` in the same formula (`cluster` is based on GEE modeling building, while `frailty` is based on mixed-effects model building).

The `summary` output for the time dependent model is: [code](#)

```
Call:
coxph(formula = Surv(patch_days, !is_censored) ~ cluster(ID) +
priv_di * (cvss_score + c_o + dis_by_s + os + y2 + smallvendor +
small_loge) - c_o - dis_by_s - os - smallvendor + cvss_score:(c_o +
dis_by_s + s_app) + opensource:c_o + opensource:dis_by_s +
os:s_app + s_app:y2, data = ISR_split)

n= 2242, number of events= 2081

          coef exp(coef)  se(coef) robust se      z
priv_di      2.798750 16.424106 0.216150 0.209360 13.368
cvss_score    0.153926 1.166404 0.016806 0.017733  8.680
y2           0.277421 1.319722 0.044042 0.044590  6.222
small_loge    0.037114 1.037811 0.007262 0.008817  4.210
priv_di:cvss_score -0.114788 0.891555 0.017327 0.016795 -6.835
priv_di:c_o     0.644347 1.904743 0.228463 0.211989  3.040
priv_di:dis_by_s 0.475405 1.608665 0.116261 0.106601  4.460
priv_di:os       -0.331847 0.717597 0.098936 0.086976 -3.815
priv_di:y2       -0.614162 0.541094 0.063296 0.061954 -9.913
priv_di:smallvendor -0.440845 0.643492 0.138900 0.099310 -4.439
priv_di:small_loge -0.082120 0.921161 0.016589 0.014449 -5.683
cvss_score:c_o    -0.060084 0.941685 0.012861 0.011990 -5.011
cvss_score:dis_by_s -0.061114 0.940716 0.008798 0.011002 -5.555
cvss_score:s_app   -0.096771 0.907764 0.014972 0.014853 -6.515
c_o:opensource     0.443978 1.558896 0.137952 0.118459  3.748
dis_by_s:opensource 0.414151 1.513086 0.091161 0.102359  4.046
os:s_app           0.815803 2.260991 0.077450 0.093536  8.722
y2:s_app           0.291007 1.337774 0.047420 0.045599  6.382

Pr(>|z|)

priv_di        < 2e-16 ***
cvss_score     < 2e-16 ***
y2            4.92e-10 ***
small_loge     2.56e-05 ***
priv_di:cvss_score 8.22e-12 ***
priv_di:c_o     0.002369 **
priv_di:dis_by_s 8.21e-06 ***
priv_di:os       0.000136 ***
priv_di:y2       < 2e-16 ***
priv_di:smallvendor 9.03e-06 ***
priv_di:small_loge 1.32e-08 ***
cvss_score:c_o    5.42e-07 ***
cvss_score:dis_by_s 2.78e-08 ***
cvss_score:s_app   7.27e-11 ***
c_o:opensource     0.000178 ***
dis_by_s:opensource 5.21e-05 ***
os:s_app           < 2e-16 ***
y2:s_app           1.75e-10 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.654  (se = 0.007 )
Rsquare= 0.231  (max possible= 1 )
Likelihood ratio test= 590.1  on 18 df,  p=0
```

```
Wald test = 376.9 on 18 df, p=0
Score (logrank) test = 586.8 on 18 df, p=0, Robust = 454.6 p=0
```

(Note: the likelihood ratio and score tests assume independence of observations within a cluster, the Wald and robust score tests do not).

There are two parts to the contribution made by `priv_di`; as a standalone variable it has a large impact, but its interactions with other variables create a large impact in the opposite direction (the model building process tries to minimise its error metric, not make it easy for us to understand what is going on).

The following is the component of the fitted equation of interest:

$$e^{priv\_di(2.8 - 0.11cvvs\_scorei + 0.64c\_o + 0.48dis\_by\_s - 0.33os - 0.61y2 - 0.44smallvendor - 0.08small\_loge)}$$

where: `priv_di` is 0/1, `cvvs_score` varies between 1.9 and 10 (mean 7), `c_o` 0/1 NA other<sup>xxx</sup> (mean 0.13), `dis_by_s` 0/1 disclosed by SecurityFocus (mean 0.38), `os` 0/1 vulnerability in O/S (mean 0.26), `y2` years since 2000 (mean 1.9), `smallvendor` 0/1 small vendor flag (mean 0.25) and `small_loge` zero for small vendors, otherwise the log of number of employees (mean 5).

Applying hand waving to average away the variables:

$$e^{priv\_di(2.8 - 0.11 \cdot 7 + 0.64 \cdot 0.13 + 0.48 \cdot 0.38 - 0.33 \cdot 0.26 - 0.61 \cdot 1.9 - 0.44 \cdot 0.25 - 0.08 \cdot 5)} \rightarrow e^{priv\_di(2.8 - 0.77 + 0.08 + 0.18 - 0.09 - 1)}$$

gives a (hand waving mean) percentage increase of  $(e^{0.49} - 1) \cdot 100 \rightarrow 63\%$ , when `priv_di` changes from zero to one. The percentage change for patches for vulnerabilities with a low `cvvs_score` is around 90% and for a high `cvvs_score` is around 13% (i.e., the patch time of vulnerabilities assigned a low priority improves a lot when they are publically disclosed, but patch time for those assigned a high priority is slightly affected).

The process of calculating the 95% confidence bounds, based on the values in the summary output, is fiddly and left to the reader.

Time dependencies can appear in various forms.

A study by Koru, El Emam, Zhang, Liu and Mathew<sup>783</sup> investigated the impact of class size, in LOC, on number of defects for ten products within the KOffice suite between April 1998 and January 2006. The following are some ways in which fault characteristics can change over time:

- the number of lines of code in files change as programs evolve. Every line of code added is a potential cause of a fault.

The changing number of lines can be handled using the approach taken for the change of vulnerability visibility status. Each change in the number of lines in a class (or whatever the unit of measurement) appears as a separate row, with the rows for each class having the same id (specified in a formula as `cluster(id)`). The following data is for Kword:

```
id,start,end,event,size,state
204,0,163,1,31,1
204,163,6372,1,29,2
204,6372,11742,0,29,2
204,11742,87259,0,32,2
```

- as existing faults are discovered the probability of discovering new faults decreases, i.e., the total number of faults is finite.

To ensure consistency across classes the data is stratified by number of faults discovered. Given the wide variation in the number of faults discovered, from 0 to more than 25 per class, the number of cases in each stratification level would be small. Koru et al divided classes into four strata, containing 0, 1-5, 6-25 and >25 faults.

Stratification variables are specified in a formula using the `strata` function, e.g., `strata(state)`.

The term *recurring event* is used to describe situations where an event can occur more than once (and so is not a terminal event).

---

<sup>xxx</sup> No information is available on what this variable represents.

The following call to coxph builds a model for the Kword data (the log of lines of code has been found to provide a good fit in other contexts):

```
LOC_mod = coxph(Surv(start, end, event) ~ log(size)+strata(state)+cluster(id),
                  data=kw_data)
```

### 39\_Koru\_cox.R TODO...

If size in LOC is a robust explanatory variable, its impact should be consistent across different component products of KOffice, with some variability due to differences in kind of program (frailty modeling...).

#### 11.11.3.3 Parametric models

The first question that needs answering when building a parametric model is the form of the hazard function.

If the hazard is monotonic, i.e., continually increasing, decreasing or staying the same, then the Weibull function is the obvious first hazard function to try.

If the hazard function increases/decreases and then reverses to decreasing/increasing the Log-normal function is a reasonable first hazard function to try.

If the hazard function exhibits other growth patterns, then a piecewise approach...

[survival/vulnerabilities/patch-param.R...](#)

*Accelerated failure time* (AFT) and proportional hazards (PH) models...

#### 11.11.4 Competing risks

When more than one possible kind of event can occur, i.e., there are multiple terminal states, a competing risk model can be used. Another way of handling this kind of data is to analyse each distinct event type separately from the other event types; data involving other events is marked as censored at the time other events occur.

The Kaplan-Meier plot for a single event, in a competing risk context, may give a misleading impression of the actual situation for events that rarely occur. The *cumulative incidence curve* (CIC) is a commonly used alternative that includes information on every event (when there is only one event,  $CIC = 1 - KM$ ). CIC does not assume that competing risks are independent and estimates the marginal probability of an event.

The cmprsk package includes support for competing risk models.

A study by Di Penta, Cerulo and Aversano<sup>349</sup> investigated the history of problems in source code flagged by various static analysis tools. Newly written source code may contain a construct that is flagged and this code is tracked through subsequent versions. Possible competing events include the removal of the code containing the flagged construct and the flagged construct being modified such that it is no longer flagged (i.e., a bug fix).

Figure 11.75 shows the cumulative incidence curves (created by the cuminc function in the cmprsk package) for problems reported in the Samba and Squid source by the splint static analysis tool.

```
library("cmprsk")

plot_cif=function(sys_str)
{
  t=cuminc(rats$faultime, rats$type, cencode=0, subset=(rats$SYSTEM == sys_str))

  plot(t, col=pal_col, cex=1.25,
        curvlab=c("was removed", "disappeared"),
        xlab="Snapshot", ylab="Proportion flagged issues 'dead'\n")

  text(max(t[[1]]$time)/1.5, 0.9, sys_str, cex=1.5)
}

plot_cif("samba")
plot_cif("squid")
```

Comparing competing risks...

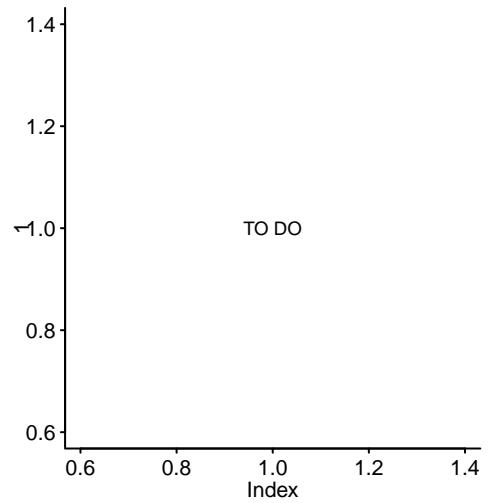


Figure 11.74: Survival curve after adjustment for explanatory variables... code

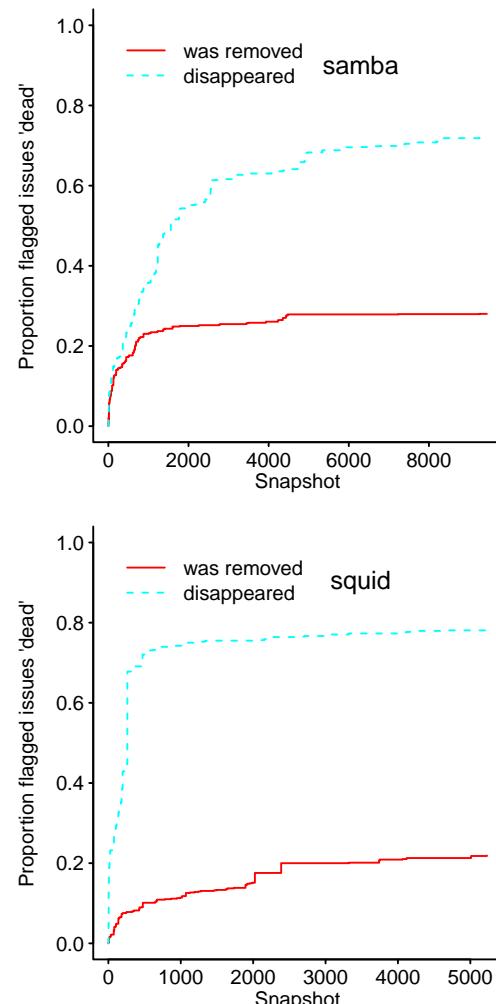


Figure 11.75: Cumulative incidence curves for problems reported by the splint tool in Samba and Squid (time is measured in number of snapshot releases). Data from Di Penta et al.<sup>349</sup> code

### 11.11.5 Multistate models

Multistate models deal with time to event processes that involve multiple events and potentially changes of state between them... TODO

The `msm` package...

?

5,855 fault records from Chinese software house... used in?

Multiple licenses that software can switch to using...?

Function creation, zero or more modifications and possible deletions. function lifetime...

## 11.12 Structural Equation Models

?

## 11.13 Circular statistics

Some measurements use a circular scale, with values wrapping back to the minimum value when incremented past the maximum value, e.g., time of day or days of the year. Circular statistics<sup>1083</sup> is the name given to the analysis of data measured using such a scale. throughout this subsection. Circular statistics has only become more widely studied in the last 40 years or so, and techniques for handling operations that are well-established in other areas of statistics are still evolving. The `circular` package is used in this section.

Differences between measurements on a circular and linear scale include the following:

- plotting uses a polar representation, rather than x/y-axis (the `circular` package includes support for the `plot`, `lines`, `points` and `curve` functions),
- the mean, if it exists, has two components: mean direction ( $\bar{\theta}$ , an angle) and mean resultant length ( $\bar{R}$ ), returned by the `mean` and `rho.circular` functions respectively (the `trigometric.moment` function provides another way of obtaining this information). The `median.circular` function returns a median (multiple medians may exist, but only one is returned),
- the term variance, on its own, is ambiguous. The *circular variance*,  $V$ , is defined as  $V = 1 - \bar{R}$  and varies between zero and one. Another measure is *angular variance* (returned by the `angular.variance` function) which varies between zero and two.  
The *circular standard deviation* is returned by the `sd.circular` function (it is not calculated by taking the square root of the variance; its formula is:  $\sqrt{-2 \log \bar{R}}$ ),
- the von Mises distribution plays a role similar to that filled by the Normal distribution on linear measurement scales.

The mean resultant length,  $\bar{R}$ , is a measure of how spread out data points are around the circle. If the points have a symmetric distribution  $\bar{R}$  equals zero and if all the points are concentrated in one direction  $\bar{R}$  equals one; for unimodal distributions the term *concentration* is applied to  $\bar{R}$  to denote the extent to which measurements concentrate around the mean direction.

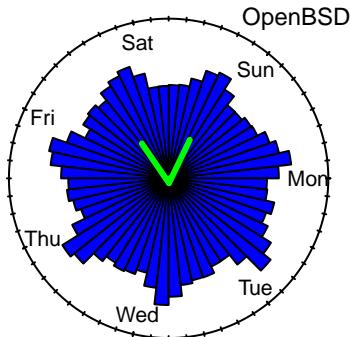
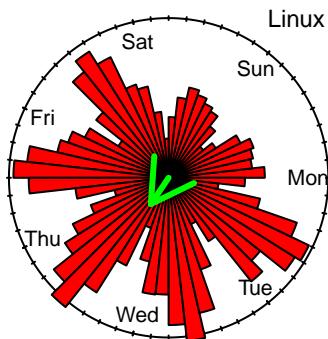
Figure 11.76 is a Rose diagram of the number of commits to Linux and FreeBSD for each 3 hour period of the days of the week (the same data is plotted using a linear scale in Figure 5.4).

The `rose.diag` function plots Rose diagrams. By default, the area of each segment is proportional to the number of measurement points in the segment (the behavior used when plotting histograms).

```
library("circular")
```

```
# Map to a 360 degree circle
HoW=circular((360/hrs_per_week)*week_hr, units="degrees", rotation="clock")
rose.diag(HoW, bins=7*8, shrink=1.2, prop=5, axes=FALSE, col=col_str)
axis.circular(at=circular(day_angle, units="degrees", rotation="clock"),
```

Figure 11.76: Rose diagram of number of commits in each 3 hour period of a day for Linux and FreeBSD. Data from Eyolfson et al.<sup>402</sup> code



```

labels=c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))

text(0.8, 1, repo_str, cex=1.4)
arrows.circular(mean(HoW), y=rho.circular(HoW), col=pal_col[2], lwd=3)

```

The arrow at the center shows the direction of the mean and the length of its shaft is its resultant length. Linux has fewer commits at weekends, compared to weekdays and a mean direction near the middle of the week looks reasonable. The number of commits to FreeBSD does not seem to vary between days; the mean length is 0.03 (it almost does not have a mean), compared to Linux's mean length of 0.2.

If the measurement scale is very coarse (e.g., measuring commit time to an accuracy of day rather than hour or minute), then  $\bar{R}$  will be underestimated and introduce errors in the calculation of the various location measures which use this value (see [statistics/circular/circle-bin.R](#)). The suggested correction to  $\bar{R}$  for calculating circular standard deviation, when measuring in units of days rather minutes, is to use the calculated value of  $\bar{R}$  multiplied by 1.034 (calculating higher order moments involves much larger values).

### 11.13.1 Circular uniformity

The *continuous circular uniform distribution* is the fundamental circular model; for this model no direction is any more likely than another. The `dcircularuniform` and `rcircularuniform` functions, but not `p` and `q` forms, are supported by the `circular` package.

The choice of circular uniformity test to use for measurements on a continuous scale, i.e., many possible measurement points around the circle, depends on how the data is thought to possibly deviate from uniformity. The two uniformity deviation possibilities are:

- a single peak over some range of values, i.e., a unimodal distribution. In this case the Rayleigh test is the most powerful known test, available in the `rayleigh.test` function,
- multiple peaks in the distribution of values around the circle. There are three tests that are more powerful than the Rayleigh test when the data distribution could be more complicated than a single peak, but no single one is superior to the others; available in the `kuiper.test`, `watson.test` and `rao.spacing.test` functions.

Unless there is a good reason to think that the measurements could have a single peak, one (or all) of the omnibus tests should be used.

When the measurements have been grouped into a few bins, e.g., months of the year, a grouped data test has to be used...

Bootstrapping [statistics/circular/grp-data-boot.R](#)...

Figure 11.77 shows how the shapes of three popular symmetrical single peak wrapped circular distributions differ from each other.<sup>xxxi</sup> The *Jones-Pewsey distribution* includes all of them, and others, as special cases.

Figure 11.78 shows asymmetric extended forms of some common circular distributions. The `circular` package does not include support for asymmetric distributions, but code is available in Pewsey et al.<sup>1083</sup>

Compiler writers born in February... TODO

A study by Eyolfson, Tan and Lam<sup>403</sup> investigated the correlation between commit time and the likelihood of a fault being detected in the commit, for Linux and PostgreSQL. Figure 11.79 shows the number of non-fault commits (upper) and number of commits in which a fault was detected (lower), made in each hour of combined weekdays (the pattern of commits on weekdays differs from weekend days and the following analysis is based on weekdays only).

What differences, if any, exist between the two sets of daily commit times and in particular are commits made at certain times of the day more likely to have a fault detected in them?

<sup>xxxi</sup> The implementation of the Cartwright distribution, up to version 0.4.7 of the `circular` package, uses the spelling `carthwrite`.

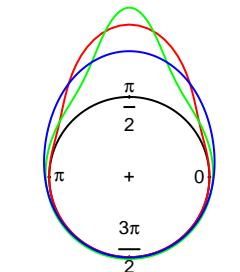
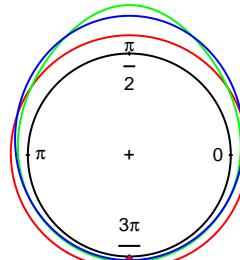
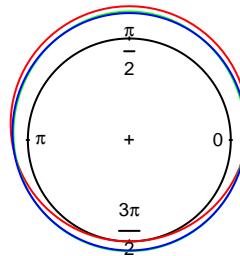


Figure 11.77: The Cartwright (red; `dcarthwrite`), wrapped Cauchy (green; `dwrappedcauchy`) and wrapped von Mises (blue; `dvonmises`) circular probability distributions for various values of their parameters. [code](#)

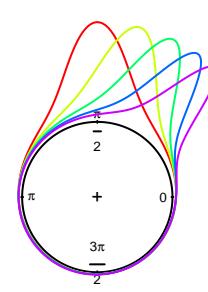
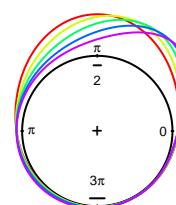
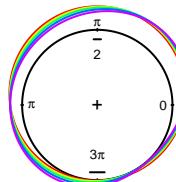
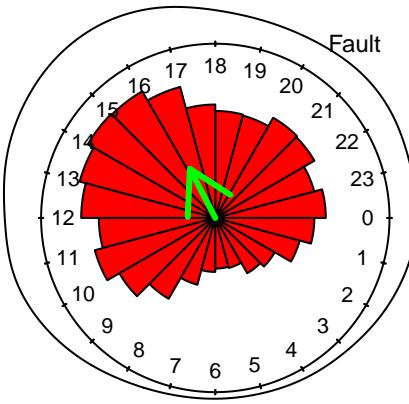


Figure 11.78: Asymmetric extended forms of the Cardioid (upper), von Mises (middle) and Cauchy (lower) probability distributions for various values of their param-



- testing for a common mean direction: The `watson.williams.test` function, in the `circular` package, assumes that both samples are drawn from a von Mises distribution; the *Watson large sample non-parametric test* does not even require the samples to share a common shape (see `statistics/circular/common-mean.R`). When any of the samples has a size less than 25, a bootstrap version of these tests should be used. The daily commit times do not share a common mean direction (15.5 hours for fault commits and 16.2 hours for non-fault commits); the mean result lengths are 0.33 and 0.32 respectively,

- testing for a common concentration: Are the points concentrated around a common direction? The *Wallraff test* is not supported by the `circular` package, but is described in Pewsey et al<sup>1083</sup> (see `statistics/circular/common-concen.R`). The two commit samples do not share a common concentration.

### 11.13.2 Fitting a regression model

When one or more variables are measured on a circular scale the technique used to build a regression model depends on whether the circular variable is an explanatory or response variable.

When the response variable is measured on a linear scale, existing techniques and functions can be used; there may be one or more circular or linear explanatory variables,

When the response variable is measured on a circular scale, the `lm.circular` function, in the `circular` package, can be used

#### 11.13.2.1 Linear response with a circular explanatory variable

Circular explanatory variables can be modeled using periodic functions and the regression modeling techniques discussed in earlier sections. The sine and cosine functions can be combined to model any periodic function. As always, a model containing the fewest number of distinct parameters is desired.

The cosine function can be modified in various ways to change its shape:

$$y = \alpha + \beta \cos(\omega x + \phi)$$

higher order harmonics can be added:

$$y = \alpha + \beta_1 \cos(\omega x + \phi) + \beta_2 \cos(2\omega x + \phi) \dots$$

The shape of the peaks and troughs can be modified by adding a sine wave to the angular argument. In the following a positive  $\lambda$  sharpens the peaks and flattens the troughs while a negative  $\lambda$  has the opposite effect.

$$y = \alpha + \beta \cos(\omega x + \phi + \lambda \sin(\omega x + \phi))$$

a skewed period (which is what asymmetrical distributions have) can be modeled by adding a cosine wave to the angular argument (provided  $-\pi/6 \leq \lambda \leq \pi/6$ , outside this range it also effects other shape characteristics):

$$y = \alpha + \beta \cos(\omega x + \phi + \lambda \cos(\omega x + \phi))$$

These are all non-linear equations and the `nls` function can be used to fit them.

The commit data, seen in Figure 11.79, is asymmetrical and the following code fits an extended cosine regression model (the `gam` values were estimated from the height of the cycle and `omega` from fitting 24 hours into  $2\pi$  radians):

```
basic_mod = nls(freq ~ gam0+gam1*cos(omega*hour-phi+nu*cos(omega*hour-phi)),
                 start=list(gam0=800, gam1=700,
                            omega=0.3, phi=1, nu=0),
                 data=week_basic)
```

The upper plot in Figure 11.80 shows the number of non-fault commits per hour for every week day and the fitted model, the lower plot shows the commits with detected faults.

Both fits handle the skewed period but not the sharp peak and flat trough. A sine contribution can be added to help handle this shape and improve the fit, the call to `nls` is below:

v 0.10.0

Figure 11.79: Number of commits (upper) and number of commits in which a fault was detected (lower) by hour of day of the commit, for Linux. Data from Eyolfson et al.<sup>403</sup>

code

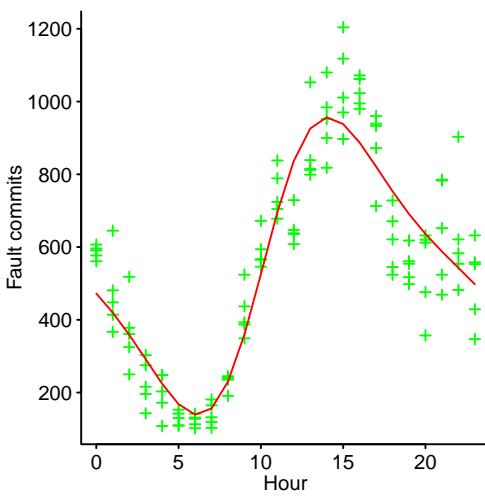
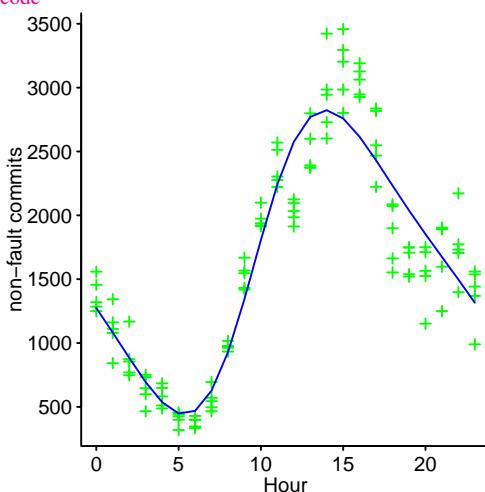


Figure 11.80: Number of commits per hour for weekdays and fitted model (upper) and number of commits in which a fault was detected (lower), for Linux. Data from Eyolfson et al.<sup>403</sup>

```
basic_2mod = nls(freq ~ gam0
+gam1*cos(omega*hour-phi+nu*cos(omega*hour-phi))
+gam2*cos(2*omega*hour-phi+nu*sin(omega*hour-phi)),
  start=list(gam0=800, gam1=700, gam2=100,
  omega=0.3, phi=1, nu=0),
  data=week_basic)
```

Figure 11.81 overlays the fitted curve for non-fault and fault (red) commits over the non-fault hourly commits for each workday.

Confidence intervals...

### 11.13.2.2 Circular response variable

The `lm.circular` function supports circular response variables... if only we had some data...

## 11.14 Compositions

Compositional data is made up of components whose total contributions sum to 100 (or 1). The requirement of a total creates a correlation between the components, i.e., if one of them increases one or more of the others has to decrease. When this mutual correlation between variables is not taken into account, models with surprising can be fitted to data.

The theory needed to underpin techniques for handling compositional data became available at the start of the century; it is all very new and many issues are still unsolved.

The analysis in this section is based on the book by Boogaart and Tolosana-Deldago.<sup>1386</sup>

A study by Machiry, Tahiliani and Naik<sup>881</sup> measured the performance of two application test generators by comparing the number of lines of program source code covered by the tests generated by each tool (50 Android apps were tested); human performance was also measured.

The application source lines covered by human and tool generated tests was recorded. The difficulty of creating tests to cover source lines is likely to vary across applications and within different parts of the same application. Normalizing coverage counts, to a percentage of source lines covered, allows performance across different applications to be compared.

Figure 11.82 shows that as the number of application source lines increases, coverage common to human and Dynodroid written tests decreases (measured as a percentage of all covered lines).

One measure of human vs. tool performance is to compare just those source lines that are covered by tests. What percentage, for each application, is covered by both human and tool generated tests and the percentage uniquely covered by human or tool tests? Figure 11.83 shows this information for the Dynodroid tool (the red tick marks on the axis are measurement points where one of the three components is zero), along with a (poorly fitting green line) regression line.

The clustering of points near the Human & Dynodroid vertex shows that tests created generated by these generators tend to cover the same source lines. More points are near the Dynodroid axis than the Human axis, suggesting that Dynodroid generated tests cover fewer unique source lines.

Fitting three regression models, one for each coverage percentage, using application source lines as the explanatory variable fails to make use of all the available information, i.e., the relationship between the three percentages.

A method of combining the three percentages into a single entity, that can be used as a response variable is required. The *isometric log-ratio transformation*, `ilr`, is one possibility and the `compositions` package supports the `ilr` function.

The `acomp` function normalises the listed columns using a ratio scale and returns an object having class `acomp` (named after Aitchison who pointed out the useful mathematical properties that a ratio scale bring to compositional analysis).

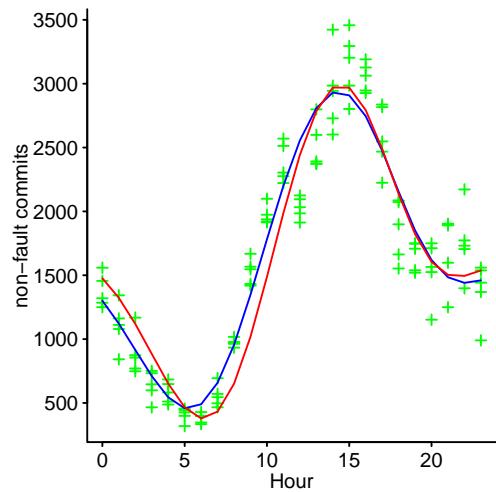
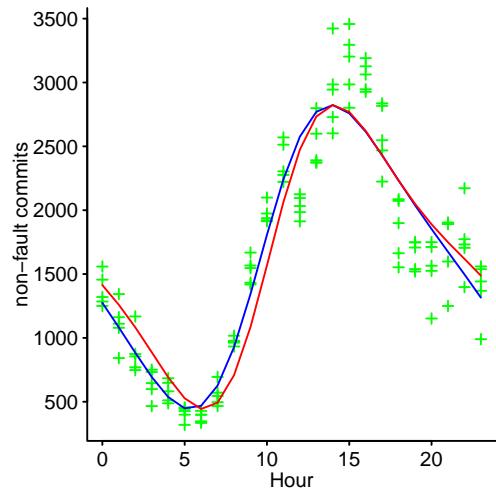


Figure 11.81: Number of commits per hour for each weekday, fitted using  $\cos(\dots\cos\dots)$  (upper) and  $\cos(\dots\cos + \sin\dots)$  (lower), for Linux; in both cases the fitted fault model (red) has been rescaled to allow comparison. Data from Eyolfson et al.<sup>403</sup> code

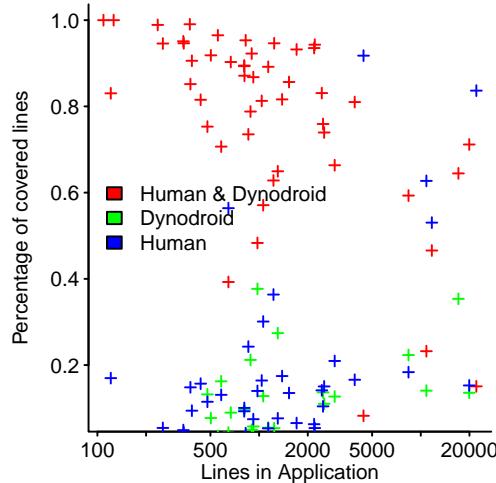
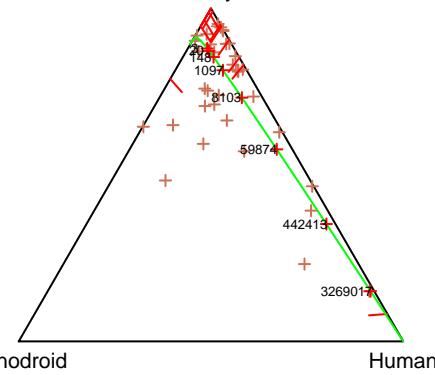


Figure 11.82: Application source lines against percentage of covered lines achieved by both Human & Dynodroid tests, only by Dynodroid tests and only by Human tests. Data from Machiry et al.<sup>881</sup> code

Human &amp; Dynodroid



The `ilr` function is not currently handled by `glm`, so `lm` has to be used. Understanding the rest of the code requires a lot more background knowledge than is appropriate here; see Boogaart and Tolosana-Delgado<sup>1386</sup> for more details.

```
library("compositions")
covered=acomp(dh, parts=c("LOC.covered.exclusively.by.Dyno..D.",
                         "LOC.covered.exclusively.by.Human..H.",
                         "LOC.covered.by.both.Dyno.and.Human..C."))
plot(covered, labels="", col=point_col, mp=NULL)
ternaryAxis(side=0, small=TRUE, aspanel=TRUE,
            Xlab="Dynodroid", Ylab="Human", Zlab="Human & Dynodroid")
dh$l_total_lines=log(dh$Total.App.LOC..T.)
comp_mod=lm(ilr(covered) ~ I(l_total_lines^2), data=dh)
d=ilrInv(coef(comp_mod)[-1, ], orig=covered)
straight(mean(covered), d, col="green")
```

The explanatory variable is total source lines in the application and the red plus signs show predictions for various totals. The quality of the fit is very poor, with potentially many outliers and non-constant variance. Model building can only use the explanatory variables present in the data.

[584](#) ...

## 11.15 Extreme value statistics

Extreme value statistics deals with values that rarely occur during the normal operation of a system... .

?  
?  
?

# Chapter 12

## Other techniques

### 12.1 Machine learning

Building a regression model requires an investment of people time and expertise, potentially a lot of time. Machine learning can often find patterns in large datasets with relatively little upfront investment of people time. The problem with models built using machine learning is often the difficulty of interpreting their behavior, i.e., they are designed for prediction performance, not ease of understanding.

The machine learning approach to model building is ideal for clueless button pushers, being clueless about a data set that needs to be analysed happens to all of us from time to time. The only input to a machine learning approach is the data,<sup>i</sup> there is no need to specify a functional form for the relationship between explanatory variables.

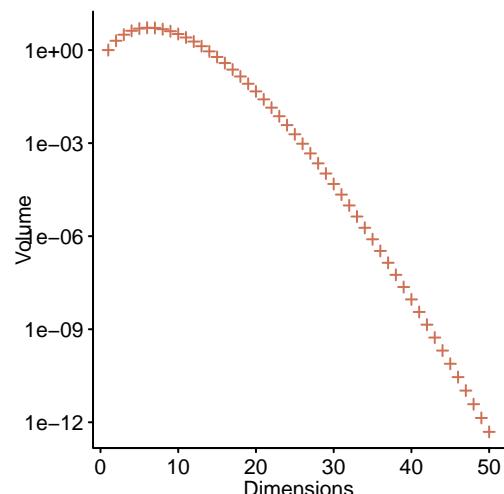
Possible uses for machine learning include:

- a filter that can quickly find the subset of important explanatory variables in a dataset containing very many variables. Once located, this subset can be fed into the regression modeling processes discussed earlier. This is essentially a fishing expedition and there will always be some subset that is better than the others. Cross validation needs to be used to check that the variables could have explanatory behavior with a different dataset,
- to build models for data whose characteristics regularly change so much that existing models become unusable.

A sample containing information about very many variables is useful to have when it is possible to use domain knowledge to select an appropriate subset. However, when all the variables are included in the analysis at the same time the *curse of dimensionality* is invoked by the fundamental mathematics often used to build models.

A common metric used by machine learning algorithms is the distance between points. Each measurement can be viewed as a point in an  $n$ -dimensional space, where  $n$  is the number of attributes associated with each measured item. For ease of comparison in the following analysis this  $n$ -dimensional space is normalised so that every side has length one, its volume is also one. In 3-dimensions the volume of a sphere of diameter one is  $\frac{4}{3}\pi 0.5^3 \rightarrow 0.52$ , that is the sphere occupies 52% of the unit cube. If the unit cube contains multiple points, then there is a 52% probability that a point at the center of the unit cube will be within 1-unit distance of another point. As the number of dimensions increases the sphere/unit cube volume ratio increases to a peak at five dimensions and then decreases rapidly. Figure 12.1 shows how the volume of a sphere changes as the number of dimensions increases.

As the number of dimensions increases the distance from a point to the point nearest to it approaches the distance to the point furthest from it;<sup>136</sup> an effect that can occur for as few as 10-15 dimensions. This behavior means that any algorithm relying on distance between points effectively ceases to work at higher dimensions.



<sup>i</sup> Implementations often support a variety of tuning parameters and the cost of ignoring them is often cpu time.

### 12.1.1 Decision trees

As the name suggests decision trees have the form of a tree like structure, rather than that of an equation. Each node of the tree contains either an expression whose result is used to select which of the node branches to follow or a value denoting the result. The commonly used rpart package supports the creation of binary trees; the Weka machine learning system<sup>1460</sup> supports nodes containing more branches and can be accessed from R using the RWeka package.

Tree models are popular in disciplines where they can be interpreted by people who need to make a decision based on what they observe, e.g., Doctors.

The predictions made by decisions tree models are generally not as accurate as other types of models, and they do not use continuous variables effectively. But these are not important issues when the aim is to gain a better understanding the data.

The condition at each node involves one variable and a set of one or more constants. The binary value of the relationship selects which branch to go down to the next node (the displayed tree goes left when the condition is true and right when false), with the process continuing until a leaf node is reached.

The rpart function decides whether a leaf node should be split into a condition node and two leaf nodes using a method known as *cost complexity pruning*; any node split that does not improve the overall fit by a factor of CP is not attempted.

A study by Shihab et al<sup>1246</sup> looked at reopened faults in the Eclipse project. Of the 18,312 bug reports 3,903 were resolved (i.e., closed at least once) and 1,530 of these could be linked to code changes. Of the 1,530 that could be linked to code changes, 246 had been reopened at the time of the study. Shihab et al cast their net very wide and extracted 22 factors, possible associated with reopened faults, from the source code repositories.

It is possible to plot a decision tree, but for non-trivial models the visualization has little practical use. Figure 12.2 shows the first few levels of the decision tree built from the Shihab et al data.

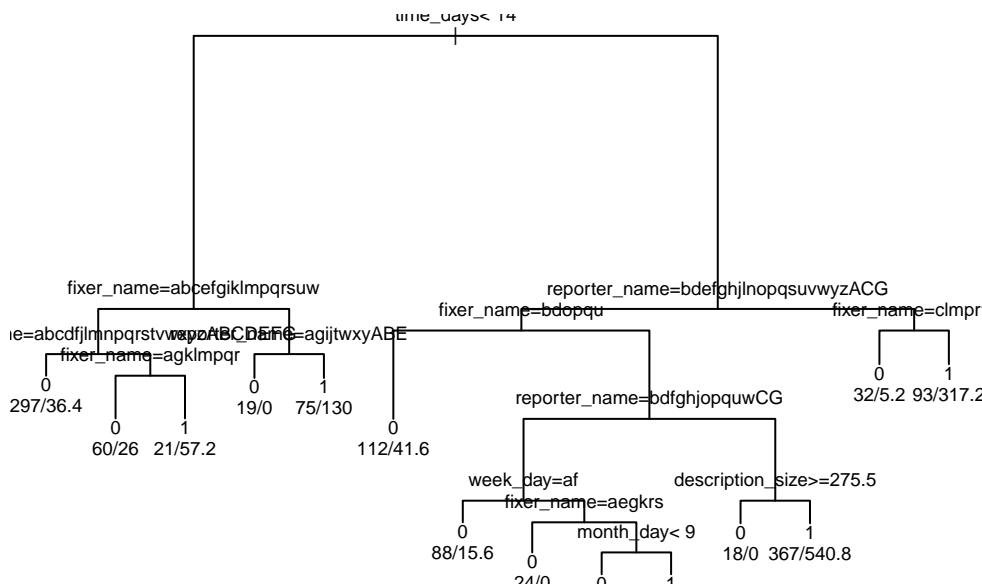


Figure 12.2: Top levels of the decision tree built from the reopened fault data. Data from Shihab et al.<sup>1246</sup> code

Which variables were found to contribute most to the model? This information can be obtained by calling summary (the cp=0.4 argument removes lots of details from the output; the printcp function does not provide any information on variable importance); the output below is for the reopened fault model:

```
> summary(weighted_model, cp=0.4)
Call:
rpart(formula = remod ~ time + week_day + month_day + month +
       time_days + severity + priority + pri_chng + num_fix_files +
       num_cc + prev_state + description_size + fixer_exp + fixer_name +
       reporter_exp + reporter_name, data = raw_data, weights = data_weight,
       method = "class", x = TRUE, parms = list(split = "information"))
```

```
n= 1530
```

	CP	nsplit	rel error	xerror	xstd
1	0.17010632	0	1.0000000	1.0697311	0.01974351
2	0.02814259	1	0.8298937	0.9340213	0.01974351
3	0.02751720	2	0.8017511	0.8742964	0.01962785
4	0.02095059	5	0.6957473	0.8656973	0.01960530
5	0.01485303	6	0.6747967	0.8813321	0.01964519
6	0.01414947	7	0.6599437	0.8642902	0.01960146
7	0.01407129	9	0.6316448	0.8355222	0.01951425
8	0.01219512	10	0.6175735	0.8302064	0.01949628
9	0.01000000	13	0.5795810	0.8345841	0.01951113

#### Variable importance

	fixer_name	reporter_name	time_days	week_day
description_size	32	32	13	6
	4	fixer_exp	reporter_exp	month_day
	priority	severity	prev_state	num_fix_files
month	1	1	1	1
	1			

Node number 1: 1530 observations

```
predicted class=0 expected loss=0.4990637 P(node) =1
  class counts: 1284 1279.2
probabilities: 0.501 0.499
```

The variables found to make a useful contribution to the model and a measure of their relative importance appears at the end (at least when a large cp argument is passed).

For the identity of people fixing and reporting problems to play such a large role in the model suggests that either this subset of people have some characteristic or the faults they close have some characteristic that causes the faults they close to be reopened. A useful pointer for further investigation.

The columns of numbers in the middle of the output contain two measures of error for various values of CP. Decision trees are susceptible to overfitting and the xerror column estimates the error using ten-fold cross validation (the error listed in the rel\_error column does not use cross validation and gives a rosier estimate). The output above suggests that building a model using the CP value listed in the second row is likely to produce more accurate results than other values (the default value of CP is 0.01).

?

[odds-and-ends/wcre2012-delaystudy.R](#) find out which variables are the biggest predictor of delayAfterChange...

## 12.2 Clustering

By dividing items in the world into categories of things, people reduce the amount of item specific information they need to learn, [1108](#) information on an item that has not been encountered before can be inferred by deducing the category it is most likely to be a member of and then applying what is known about the chosen category; studies have found that people are sensitive to the costs and benefits of using categories [886](#)...

Many clustering algorithms will always produce a clustering of the data; the clustered returned may just be patterns that happen to exist in random data. The VAT and iVAT functions, Visual Assessment of (Clustering) Tendency and the improved version, in the seriation package can be used to obtain some idea about the number of clusters that may be present in data...

?

Sequence mining...?

?

Archetypal analysis describes individual data points based on the distance from extreme points, whereas cluster analysis focuses on describing its segments using the average members as the prototypes. The [archetypes] rpackage...?

Identifiers splitting/expansion by lots of subjects.<sup>561</sup> Are there clusters of developers making similar choices?...

Cluster of log file messages... Given the diverse nature of event log entries, often undocumented, the first step is to obtain a list of the message types present in the log file.

In the `plot` there appears to be vertical banding in the plot and a horizontal grouping of very popular installations...

languages used together, known by same person, emailed...?

?, ?, ?

Building a phylogenetic tree...?

?

### 12.2.1 Principal component analysis

Reduces the dimensionality of a multivariate dataset...

The principal components are transformed linear combinations of existing explanatory variables, these principal components are uncorrelated with each other...

Used when there are too large number of explanatory variables or when the explanatory variables are highly correlated with each other...

Primarily an exploratory technique, but... Sometimes the principal components are treated as an end in themselves and are then interpreted in the same way that explanatory variables might be *explanatory factor analysis*...

The biplot function...

Independent component analysis ...

Some people claim factor analysis is not worth the time needed to understand it...

### 12.2.2 Seriation

Seriation involves finding a linear order for items that minimises/maximises some metric, with the hope that the linear order obtained has a structure that can be interpreted in a meaningful way.

The `seriation` package includes support for a range of algorithms that attempt to find some kind of optimal linear ordering of items. Evaluating all possible item orderings is impractical for all but the smallest samples (the number of possibilities grows as  $n!$ ) and so heuristic algorithms are used.

A study by Jones<sup>703</sup> investigated the extent to which developers create similar data structures to hold information listed in a specification, i.e., grouping together identifiers containing related information in the same data structure. The hypothesis was that shared cultural and professional experiences would cause subjects to define similar data structures.

Subjects were given a list of items from the “Department of Agriculture” and asked to design a C/C++ API containing this information. The results list, for each subject, the structs or classes defined and their fields/members (with each field containing one item of API information, e.g., “Date crop harvested” and “Organically produced”).

Depending on the question asked and the form of the data various techniques are available.

Figure 12.3 shows which items are placed in the same data structure as one particular item, in this case “Antibiotics used”, by each subject. The matrix passed to `seriate` contained boolean values (in the same data structure or not) and subjects/fields are ordered.

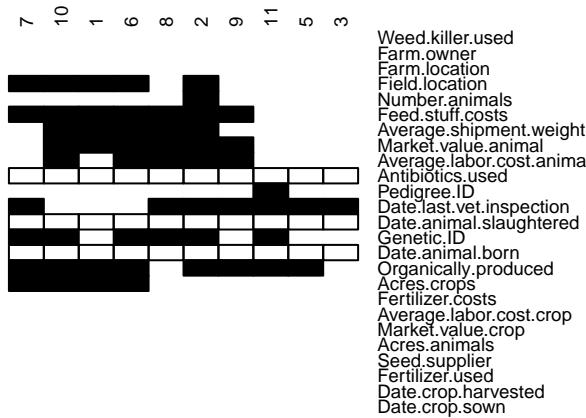
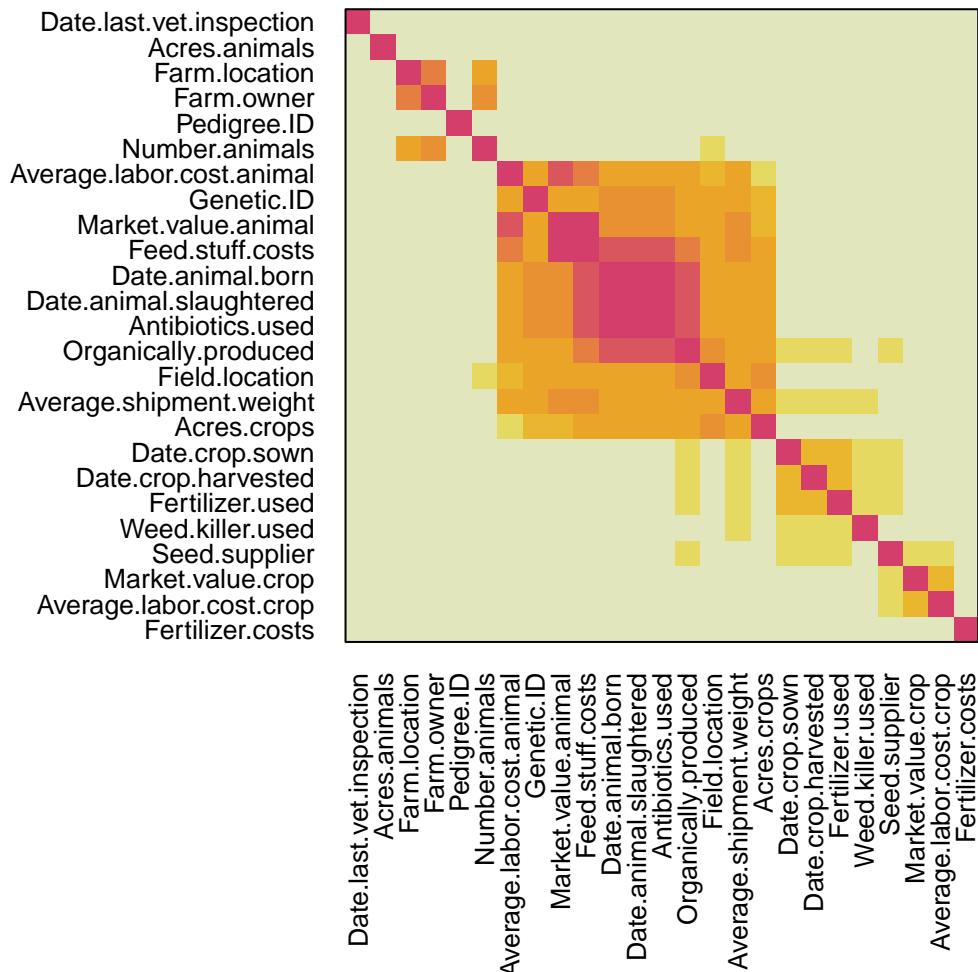


Figure 12.3: A Bertin plot for items included in the same data structure as “Antibiotics used”, for each subject, after reordering by seriate. Data from Jones.<sup>703</sup> [code](#)

```
library("seriation")
fser=seriate(fmat, method="BEA", control = list(rep = 10))
bertinplot(fmat, fser, options=list(panel=panel.squares, spacing=0,
gp_labels=gpar(cex=0.6)))
```

A single item’s pattern of association with all the other items can be generalised by counting the number of times every pair of items occurs together in the same data structure. A Robinson matrix has the property that the value of its matrix elements decrease, or stay the same, when moving away from the major diagonal and this matrix has been used to study commonality in subjects’ categorization behavior.<sup>1176</sup> Figure 12.4 shows a visualization of a Robinson matrix.



```
library("seriation")
fdist = as.dist(1 - fmat/max(fmat)) # Normalise counts
```

Figure 12.4: A visualization of the Robinson matrix based on number of times pairs of items co-occur in the same data structure (the closer to the diagonal the more often they occur together). Data from Jones.<sup>703</sup> [code](#)

```
fser = seriate(fdist, method="BBURCG")
pimage(fdist, fser, col=pal_col, key=FALSE, gp=gpar(cex=0.8))
```

## 12.3 Ordering of items

A list of items may have a preferred ordering. The ordering may be the result of rankings, rating or paired a comparison between pairs of items. Techniques are available to calculate a preference scale for the ordering of items.

Bradley-Terry statistics are the traditional method of ordering a list of items based on results from pairwise comparisons, e.g., football match results. This method has been extended to make use of results from comparing more than two items. This section is based around use of the BradleyTerry2 package for simple paired comparisons and the prefmod<sup>585</sup> package for the more complicated cases.

[hyper2](#)<sup>578</sup> package...

An alternative approach to analyzing item ordering is discussed in Section 9.6.1.

A study by Jones<sup>700</sup> investigated developer beliefs about binary operator precedence. Subjects saw an expression, such as  $a + b \% c$ , and were asked to insert parenthesis such that the behavior (as interpreted by the compiler) remained unchanged. Bradley-Terry statistics can be used to calculate the relative precedence of the binary operators used in the study, based on subject answers (which may be different from the actual precedence).

The parenthesis indicates the operator which can be considered to *win* the precedence choice. The BTm function, in the BradleyTerry2 package<sup>1370</sup> takes the results from paired comparisons, and returns the coefficients of a fitted model (internally, the glm.fit function is used to fit a logit regression model); the estimated coefficients denote differences in operator precedence

```
library("BradleyTerry2")
prec_BT=BTm(win_win, first_op, second_op, data=nodraws)
summary(prec_BT)
```

The output returned by the summary function has the same form as that returned by other regression models. Note: the summary function does not list the factor with value zero (the subtraction operator, for this data). The BTabilities function lists all factors, and their values.

[sourcecode/jones\\_prec.R](#)...

In a sports match the home team is often considered to have an advantage over the away team. Perhaps, when developers are uncertain they are more likely to select the first binary operator, of a pair. A model can be updated to include information on first position *wins*, for each operator pair (the effect is very small); see [sourcecode/jones\\_prec.R](#) for details.

Three or more items can be compared by ranking them in order of preference. The techniques used to analyze paired comparisons can be extended to include ranked preferences; the pattR.fit function in the prefmod package can be used to analyse ranked data (and the patt.worth function to map it to a plotable form).

A study by Biegel, Beck, Hornig and Diehl<sup>138</sup> investigated the ordering of definitions appearing in 5,372 classss, from 16 Java programs. The Java coding conventions (JCC)<sup>1326</sup> recommends a particular ordering of the four kinds of definitions, and on average over 80% of classes in these projects follow the JCC recommendation: *class variable* (or *field*), *instance variable* (or *static initializer*), *constructor* and *method*.

Definitions include access control information, via the use of the keywords: private, protected, public and no keyword (the default behavior given in the language specification is used). The ordering of definitions by identifier visibility (i.e., access control) can be treated as a ranking.

Figure 12.5 shows the ordering (ranking) of definitions by access control keyword, over all projects investigated; see code for details (most of the code is needed to convert the data to an array of rankings, that is passed to the `pattR.fit` function).

The y-axis values, labelled "Worth estimate", can be used to estimate the probability that an identifier defined with visibility  $V_i$  appears before one with visibility  $V_j$  using the formula:  $\frac{V_i}{V_i + V_j}$ .

The RankAggreg package contains a variety of functions for aggregating a collection rankings to calculate a group ranking. These functions use the relative order of items in a vector to denote rank...

Lots of readability ratings...?

## 12.4 Simulation

Building a model that simulates the interaction between the agents involved in a process can be used to validate... understand...

- system dynamics: a system is represented by causal loops between all the interacting components, essentially an analogy representation of differential equations. System dynamics has been used to simulate project staffing<sup>198</sup>...
- discrete event simulation: a system is modeled in terms of the discrete events that can occur... `simmer` package...
- agent-based modeling: a set of agents having specified attributes interact with each other in a computer simulation. After a given number of time steps the state of the system is measured, with the results from many simulation runs combined to calculate probabilities for the various end-states. NetLogo is a popular system for agent-based modeling; the `RNetLogo` package provides an interface.

Staff scheduling... [projects/impl-sim.R](#)

Bass diffusion models...?

## 12.5 Text analysis

Are there differences in the descriptions and comments associated with reopened and non-reopened faults that can be used to tell them apart using...

For an example of using the `tm` package to extract and process (e.g., remove punctuation, stop words and stem words) the words in each of the fault descriptions and comments, see [faults/reopened\\_text.R](#).

Bayes rule tells us:

$$P(S|W) = \frac{P(W|S)P(S)}{P(W|S)P(S) + P(W|H)P(H)}$$

where:  $S$  is reopened,  $H$  non-reopened.

To combine the probabilities for each word associated with a given fault description/comment the naive Bayes classifier formula is:

$$P = \frac{psw_1 psw_2 \cdots psw_n}{psw_1 psw_2 \cdots psw_n + (1 - psw_1)(1 - psw_2) \cdots (1 - psw_n)}$$

Submitter correlation with description... Fixer correlation with comments...

Sentiment analysis is a popular technique for text by positive and negative opinions.

??

Before email contents can be analysed using R they need to be converted from the mbox format used by many email archives; the `tm.plugin.mail` package provides the necessary primitives. A variety of text cleaning/normalization operations should be performed, ideally including all the following:

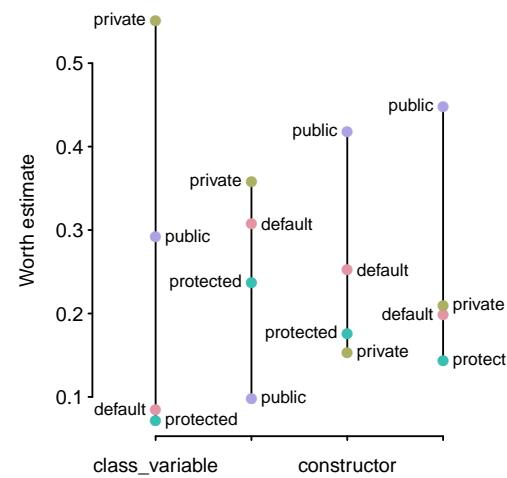


Figure 12.5: "Worth estimate" of identifier access control (visibility), for the four kinds of definitions that can appear within a Java class. Data from Biegel et al.<sup>138</sup> [code](#)

- developers sometimes use multiple email addresses within the same discussion group (e.g., because of change of job). It is important to correctly associate every email with a unique individual and so it is necessary to use some form of fuzzy matching to detect when different email addresses belong to the same person...
- the same concept may be denoted using a variety of different forms of a word. To prevent these differences having a significant impact on the detection of topic clusters the words in an email need to be reduced to a canonical form, the following are two ...
  - abbreviations: including text speak...
  - spelling: alternative (e.g., center/centre in US/UK English) and incorrect (e.g., centa)
  - stemming: a word might be written in the past tense, plural or some other way and the process of converting these different forms to a base form is known as *stemming*. Porter stemmer, Snowball...
  - synonyms: Different people will sometimes use similar, but not the same, words to denote the same concept. When a word has one or more synonyms all instances of that particular set of synonyms need to be replaced by one representative instance...

A study by Bohn, Feinerer, Hornik and Mair<sup>154</sup> analysed two R related developer discussion lists, a general help list, **r-help**, and one for those developers involved in supporting the R implementation, **r-devel**. They created the **snatm** package and the following discussion draws heavily on this...

The Bohn et al analysis used emails from the period ????, while the following analysis uses emails from the period ????.

The email transformation and cleaning consists of the following steps:

- transform the emails' format using the `makeforest` function from **snatm**, which uses the primitives provided by `tm.plugin.mail` to build a dataframe containing `rindent[emailID]`, `rindent[threadID]`, `rindent[author]`, `rindent[subjects]` and `rindent[content]` for each email.
- emails from different addresses that look like they refer to one individual are detected and normalised using the `find.aliases` function from the **snatm** package,
- normalise words using the following operations:
  - map words to lower case,
  - map words to British English using the `prepare.text` function from **snatm**,
  - replace words having one or more synonyms by a single representative word using the `wn.replace` function from the **wordnet** package (this function includes hard-coded information about Wordnet and does not replace words that are considered to be R specific; use with care on other projects),
  - remove words containing less than three characters and any that occur less than a given number of times within all emails (10 for the email Subject line and 20 for the email body),
  - remove word stems using the `stemDocument` function from the **tm** package.
- Digit sequences are removed.

A study by Jongeling, Datta and Serebrenik<sup>707</sup> found that, when applied to different issue tracking datasets, different sentiment analysis tools assigned different labels and did not agree with the manually labels assigned to a software engineering dataset.... Labeled software engineering datasets are starting to appear...?

???

?

??

# Chapter 13

# Experiments

## 13.1 Introduction

Does doing X have a significant effect on S? Traditionally X might have been a new kind of fertiliser (or drug) and S the crop yield (or being cured of some illness). In software engineering the effect sought is often a performance improvement and X the latest snake oil.

In an observational study the researcher is a passive observer, simply recording what happened or is happening; in an experimental study the researcher actively attempts to control the values of the explanatory variables (a common technique is to vary the values of one explanatory variable while the others are held constant).

A controlled experiment is the technique used to obtain the data needed to test a hypothesis. The controlled experiment that most developers are likely to be familiar with is benchmarking.

Dramatic changes in performance, after doing X, are relatively common in software development and in such cases using statistics to confirm that a noticeable change has occurred is almost a formality. At the other extreme, if a difference requires statistical analysis to be detected, it might not be a difference worth being concerned about.

Advice for running experiments often mimics the waterfall model of software development, with lots of planning and little or no feedback from production use until almost the end of the process. This advice has its roots in the environment in which experiments are carried out by the audience of many statistics text books, where running an experiment is costly (in money or time) or is a once only opportunity.<sup>i</sup>

Some questions in software engineering are amenable to iteration. Running quick, inexpensive experiments can be an efficient technique for filtering possible issues of interest and obtaining information on which of the myriad of variables have a large impact on the response of interest.

Finding the right question to ask is sometimes the most useful output from running an experiment.

An important point to remember is, that, it is better to have an inexact answer to the right question than an exact answer to the wrong question.

Like all software development activities, experiments have to pay their way. Some of the answers needed for a cost-benefit analysis include the following:

- the cost of running an experiment capable of producing the information of interest within acceptable confidence intervals,
- the usefulness of information likely to be obtained by running an experiment using a given amount of resources (such as time and money)

---

<sup>i</sup> Experiments in the social sciences, major producer of experimental studies, are often grant funded, with limited opportunities for rerunning experiments that failed to produce data that can be published.

Many software engineering tasks are performed within complex environments. Controlling and measuring all the variables in the environment is so time-consuming and expensive that few researchers are willing to attempt controlled experiments. Consequently, much of the hypothesis testing performed in commercial environments is based on convenience samples, obtained from experimentally uncontrolled, production software projects.

Experimenters want to use as many subjects as possible, because the impact of random effects on the final results usually decreases as the number of measurements increases. However, it is rare to have the luxury having access to more subjects than it is possible to make practical use. Obtaining experimental subjects often involves being forced to make do with what's available.

Software engineering is not known as an academic subject where experiments are commonly performed. A study<sup>1263</sup> of 5,453 papers in software engineering journals published between 1993 and 2002 found that only 1.9% reported controlled experiments (of which 72.6% used students only as subjects) and the statistical power of many of these experiments fell below expected norms.<sup>369</sup>

### 13.1.1 Measurement uncertainty

The term *measurement error* is often applied to measurements involving physical quantities. It is assumed that any difference between the measured and actual value is caused by errors made in the measurement process.

In software engineering, some quantities can be measured exactly, e.g., lines of code in a given program. However, the process that generated the code (e.g., a software developer) may produce a different number might a different number of lines, if it repeated using a different developer or even the original developer (reimplementing the program); see Figure 5.18.

Goodhart's law (which is an observation about human behavior, rather than a law) is "Any observed statistical regularity will tend to collapse once pressure is placed on it for control purposes." If the measurements collected were actively used to control or evaluate the development team, then the developers would be motivated to move the measurements in a direction favorable to them.

A study by Perry, Staudenmayer and Votta<sup>1078</sup> investigated differences in the measurements of various software development activities (e.g., working time on an activity and activities performed during the day), as measured by the developers involved (i.e., self reports) and as measured by external observers. Differences included developers reporting activity time that was 2.8% greater, on average, than reported by an observer, and an agreement rate on activities performed that varied between 0.6 and 0.95...

A series of studies<sup>750</sup> of social network data, as reported by those within the network and extracted from externally observed information, found that differences were large enough to render invalid any analysis of a network characteristics based on member supplied information.

The issue of random variations in performance between what are intended to be identical hardware components is discussed in Section 13.4.2.1...

## 13.2 Design of experiments

Randomization is the foundation of any claims of causation involving experimental results, i.e., doing X caused Y might be considered a valid interpretation of the data. Without randomization the most that can be said is that there is a correlation between doing X and Y occurring.

An experiment measures the performance of subjects carrying out a task in a particular environment. If the information of interest is subject performance carrying out this task in the environment used in the experiment, then the results are likely to contain exactly the information of sought. However, in practice there is not always one-to-one mapping between the experiment and practice, with differences including:

- the subjects are a convenience sample of the population of interest,

- the task used does not share all the important characteristics of the tasks that is performed outside of the experiment, or those that are shared are in very different proportions,
- the environment in which the experiment is performed is different from the one likely to exist outside of the experiment (e.g., available time may be shorter in an experiment).

A critical component of experiment design is controlling all variables that could have a significant impact on the output. Failure to take into account and control variables having a significant impact can cause a tiny effect to appear to be a large effect and vice versa. One person's tongue-in cheek-advice on how to bias an experiment to obtain the desired outcome<sup>954</sup> is another person's list of thoughtless mistakes.

Ideally all the factors that could have a significant impact on the outcome of an experiment are controlled and when they cannot be controlled their impact is contained.

One technique for handling the problem of uncontrolled variables is to group subjects into blocks based on the variable that is suspected of influencing the response (a process known as *blocking*), randomization of subjects then occurs within each block. The identity of the block becomes another explanatory variable during analysis of the results.

Subject characteristics can sometimes interfere with good experimental design, e.g., human subjects have a memory of their previous experiences that they cannot choose to erase.

A study by Basili, Green, Laitenberger, Lanubile, Shull, Sørumgård and Zelkowitz<sup>100</sup> compared the performance of perspective based reading (which instructs reviewers to read a document from a specified perspective, e.g., a designer, tester or user) against the reading technique currently used by the professional developers who were the subjects.

The researchers thought it likely that training subjects to use the new technique would change their performance on whatever technique that currently use, and decided to measure subject performance when using their current technique first, before giving them any training in the new, perspective based reading, technique.

Being forced to have all subjects use the same techniques in the same order means it is not possible to separate ordering effects in the results, e.g., learning during the experiment and any random distraction effects that only occurred at certain times.

Other factors outside of the control of the researchers, that could affect the results, include:

- the time taken for people to become proficient at using a new technique, old habits die hard. How much practice do subjects need, to reliable estimate the performance of a new technique? In this study subjects were taught PBR two days after the first part of the experiment, trained on a test document, reviewed one document, received more training and then reviewed another document.
- the kind of review technique used by subjects in the first half of the experiment. A change in performance is expected, but the details of what the change is relative to are not known (it is assumed that adhoc techniques are being used),
- the characteristics of the seeded faults. Were more faults found in the NASA documents because readers were familiar with reading that kind of document, or perhaps the characteristics of the seeded faults was such that they were harder to detect in one kind of document than another?

The experimental output included, for each subject, the number of faults detected (which has a known upper limit and a yes/no detection status) and the number of false positives (which has no upper limit, in theory) in each document reviewed. These characteristics of the data select for a binomial distribution for faults found and a Poisson distribution for the false positive distribution in the respective regression models. See [faults/basili/pbr-experiment.R...more](#)

Basing all experimental choices on random selection does not automatically create samples that maximise the information that can be obtained.

A study by Porter, Siy, Mockus and Votta<sup>1104</sup> investigated software inspections. The structure of the inspection process was manipulated by varying the number of reviewers (1, 2 or 4), number of meeting (1 or 2) and for multiple meetings whether reported faults were repaired between meetings (88 inspections occurred, involving 130 meetings and 17 reviewers).

Selecting the treatment to use, for a review, from successive entries on a randomised list of all possible treatment structure combinations (created at the start of the study) would ensure that the results contain data that is balanced across the variables of interest. However, the choice of treatment to use was randomly selected from all possibilities as each unit of code became available for review, resulting in some combinations of reviewers/meetings/repaired not being used and some used very often. The sample contained an unbalanced set of experimental conditions, making it difficult to reliably fit a model (see [experiment/porter-siy/inspection.R](#), [experiment/porter-siy/meeting.R](#) and Figure 11.32).

### 13.2.1 Subjects

Experimental subjects might be people or artefacts such as computers. An essential requirement for generalising the results from an experiment to a larger population of subjects is that the applicable characteristics of the experimental subjects are representative of the population of interest.

Typically, a limiting factor, when designing an experiment, is the amount of time subjects are likely to be willing to make available to participate.<sup>[1262](#)</sup>

There are *human factors* involved when people are the subjects in software engineering experiments; developers are intelligent beings who are constantly adapting to their environment, including the environment of an experiment (e.g., they learn and retain memories of their experiences; see «\_learning\_and\_experience»), they also experience fatigue, and attention ebbs and flows during an experiment

The major issues involved in treating computer hardware as subjects are covered in Section 13.4, while the major issues in human cognitive performance are covered in Chapter 2.

Professional developers working in different ecosystems probably share a set of basic skills and knowledge, such as being able to fluently use at least one programming language.

Much of the published research involving human subjects in software engineering experiments has used students as subjects. Students are a convenience sample for many researchers and results based on student subjects are unlikely to be questioned, i.e., it is not an issue in getting the research published. However, the results from experiments using student subjects is unlikely to be applicable to professional software developers, reasons for this include:

- students' commercial software skills and knowledge is likely to be very poor in comparison to professional developers.<sup>[978](#)</sup> This lack of experience and know-how means that student subjects will have to spend time on activities that are second nature to professionals, or they simply make noncommercial judgement calls,
- students, typically, have very little experience of writing software, perhaps 50 to 150 hours (and many have no basic coding skills<sup>[856,914,1381](#)</sup>), while commercial software developers are likely to have between 1,000 to 10,000 hours of experience. This lack of programming fluency means that student programming performance is likely to contain the effects of a strong earning component, as well as student performance being much lower than professional developers,<sup>[966](#)</sup>

Industry is well aware that students' software engineering skills are not representative of professional developers; industry is where many graduates find employment after graduation and the abilities of these new employees is plain for everyone within industry to see.

In other areas of research students subjects may be more representative of the target population, because they have had many years of experience performing the activities and tasks used in those areas, e.g., processing text written in English and everyday image processing are activities used in cognitive psychology experiments.

When experimental results are intended to be applied to the population of university students studying a computing related subject, subjects drawn from this population can be representative of it.

In the US and UK students pay to attend university and like all businesses universities have to respond to customer demand. When a student decides to study a computing related subject, perhaps because they believe it will improve their job prospects, the University's interests are in ensuring that the student meets its minimum entry requirements and can pay; the likelihood of that person being offered employment in a software related job is not a consideration.

Given the high failure rate for programming courses,<sup>3</sup> many students on such courses may not even have any software development skill or ability...

Note on terminology: many academic studies use the phrase *expert* to describe subjects who are final-year undergraduates or graduate students, with the term *novice* used to describe first-year undergraduates. In a commercial software development environment a recent graduate is considered to be a *novice* developer, while somebody with five or more years of commercial development experience might know enough to be called an *expert*.

Amazon's Mechanical Turk is becoming popular as a resource for finding subjects and running experiments (in 2015 the population of workers was estimated to be 7,300<sup>1315</sup>). Subjects can stop taking part in a MTurk experiment at any time and care needs to be taken to ensure that the characteristics of subjects who remain does not bias the results.<sup>1495</sup>

### 13.2.2 The task

A requirement for generalising the results from an experiment to the tasks performed under work conditions is that the characteristics of the experimental tasks performed by subjects share the same important characteristics as the work tasks. That is, the task needs to mimic realistic activities (the technical term is being *ecologically valid*), the two important factors are:

- being representative of real world intended usage requires obtaining reliable information about how a system will be used in real life and the inputs it is likely to experience; lack of resources to perform an analysis of real world usage often means that a convenience sample is used. In a rapidly changing environment it may not even be possible to specify usage patterns in sufficient detail and perhaps one of the important real world behaviors that needs to be benchmarked is adaptability to change.

A study by Gregg and Hazelwood<sup>549</sup> provides an example where data usage characteristics are the deciding factor in the cost/benefit trade-off. They measured the time taken to perform a matrix multiply when the CPU uses a local GPU (the SGEMM implementation in the nVidia CUBLAS package<sup>1025</sup> was used). Figure 13.1 shows that moving data between the CPU and GPU (a Barracuda12 containing a GTX 480 with 1024MB) consumes a significant amount of time relative to the work done on the data once inside the GPU. Deciding whether GPU usage is worthwhile depends on the size of matrices encountered in the real world use case, the performance may be slower because of the data transfer overhead, or faster if the matrices are large to consume the largest amount of compute resources.

Another example of the impact of variations in the input data is related to Figure 10.6.

- using a product that is identical to the one that will be used in production. Products that appear to be very similar often have very different performance characteristics (this is one reason why they exist as different products; the other common reason is marketing).

In a study by Bird,<sup>143</sup> a performance optimization expert took the existing generic code of a library and created tuned versions for each of five different processors (IBM's Blue Gene P and four different members of Intel's x86 product line). The performance of the generic and all tuned versions of the code was measured on all processors. Figure 13.2 shows relative performance, with the x-axis listing the processor the code was tuned for and the y-axis the processor on which it was run; the performance impact of executing code tuned for one processor on a different processor is dramatic.

In practice, availability of resources is often a constraining factor; for example, benchmarking backup/restore tools or desktop search applications requires that realistic file system contents be used (e.g., the file system must contain a realistic number of files, directory depth, disk fragmentation, etc); getting to a position of being able to generate realistic file systems is a non-trivial task,<sup>12</sup> let alone realistic file content characteristics.<sup>1339</sup>

### 13.2.3 What is actually being measured?

Subjects may not solve the problems they are presented with, in an experimental context, in ways that were intended by the person who designed the experiment.

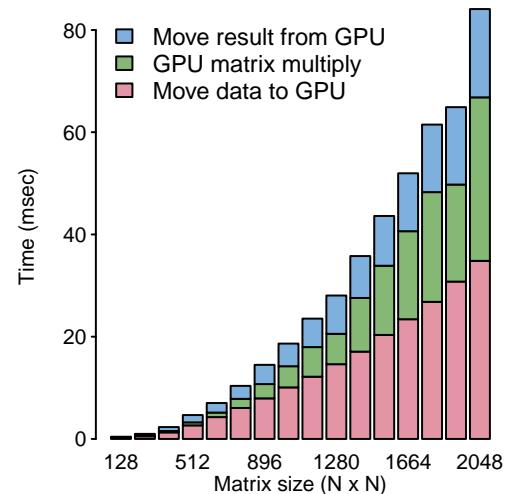


Figure 13.1: Time taken to transfer and multiply 2-dimensional matrices of various sizes on a GTX 480 GPU. Data kindly supplied by Gregg and Hazelwood.<sup>549</sup> [code](#)

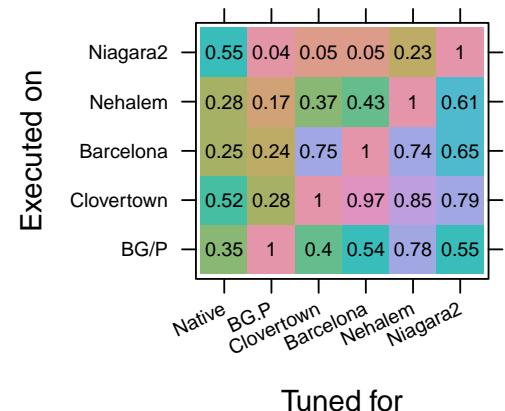


Figure 13.2: Relative performance (y-axis) of libraries optimized to run on various processors (x-axis). Data from Bird.<sup>143</sup> [code](#) July 31, 2018

Subject motivation is an important factor in obtaining reliable experimental data. Subjects who feel they are being coerced may respond by providing spurious responses or simply attempt to minimise the time spent taking part in the experiment, without attracting attention by making too many errors.<sup>590</sup>

The history of research into human memory provides an example of how early experimental results were misinterpreted.<sup>698</sup> Experiments asked subjects to remember sequence of digits and the results suggested that STM has a capacity limit of  $7 \pm 2$  items.<sup>950</sup> The  $7 \pm 2$  digit limit model was later replaced by a model based on a limit of 2 seconds of sound<sup>81</sup> (in English this corresponds to around 7 digits, 5.8 in Welsh<sup>385</sup> and around 10 digits in Chinese;<sup>637</sup> the number of digits that can be held in memory when people use these languages).

Software developers are problem solvers and get plenty of practice in finding patterns that can be used to achieve a goal. Unless an experiment is carefully constructed, it would be naive to assume that developers will use any of the techniques anticipated by the person who designed the experiment.

Your author once ran several experiments<sup>700</sup> expected to find a *2 seconds of sound* effect in developers short term memory of source code (sequences of simple assignment statements were used). A great deal of attention went into creating code sequences whose spoken form required either more or less than 2 seconds of sound, but the results did not contain any evidence for the expected effect (i.e., a difference in performance caused by the length of sound in the spoken form of source code statements).

At the end of one experiment a subject mentioned a strategy he had used to help improve his performance, remembering the first letter of each variable (I had not noticed that the variables in each list had unique first letters). Use of this strategy reduced the amount of STM the subject needed to use and is one explanation for why the expected effect was not found.

The last task on subsequent experiments asked subjects to list any strategies they used during the experiment.

A study<sup>561</sup> of how subjects source split code identifiers into components and expanded them, or not, into words measured individual performance against what was considered to be the definitive expansion of each identifier. The results of this experiment could be used to measure the performance of the researchers in creating a list of identifier expansions that maximised the likelihood of developers correctly decoding the intended information.

Source code is full of cases where what might be thought to be small differences in semantics have a large difference in usage. For instance, many languages allow numeric literals to be specified using decimal or hexadecimal notation and there is a large difference in the distribution of occurrence of literal values written using each notation; see Figure 13.3.

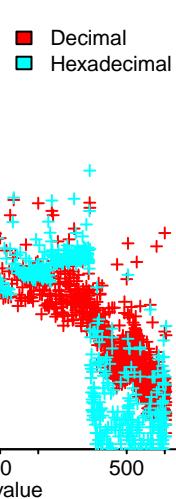


Figure 13.3: Number of integer constants having the lexical form of a decimal-constant (the literal 0 is also included in this set) and hexadecimal-constant that have a given value. Data from Jones.<sup>699</sup> code

### 13.2.4 Stopping conditions

The cost of running an experiment means it can be very tempting to stop as soon as what is thought to be a reliable result is obtained. For instance, a researcher may process subjects in batches, running statistical tests after each batch of results becomes available (e.g., checking the p-value), to check progress.

Differences in subject performance, as an experiment progresses, will cause variations in the result of statistical tests and it is possible that some cut-off value (e.g., a p-value significance level) is temporarily achieved before later result cause it to revert to a less extreme value.

Terminating the testing of subjects before results from all the preplanned subjects are available...

In an experiment terminates before completing all the planned cases is an important material fact that needs to be reported, so readers can make their own assessment of its impact on their confidence in the results.

Writeups must include a list of all the variables collected in an experiment. It does not include information on variables that did not make it into the final model, however readers need to be able to judge whether variables have been cherry picked... bonferroni correction has been applied...

Examples of ways in which A/B testing on the web can go wrong<sup>778</sup> ...

?

Sequential adaptive design for modifying an experiment as results become available... sequential sampling...

The Sequential and AGSTest packages...

Sequential probability ratio test SPRT package...

?

Using Instance selection and rough set theory to decide which options to use for the next sample...?

### 13.2.5 Selecting experimental options

The behavior of many systems depends on the setting of a wide variety of options and an estimate of the impact of individual options on system performance may be required. One way of obtaining this estimate is to measure system performance for all possible combinations of option values. This approach might be practical for a small numbers of options, each having relatively few values, e.g., Apache supports nine build time yes/no options giving  $2^9$  possible configurations (out of these 512 only 192 are valid). However, many large systems often support some many options that building and executing every configuration would be impractical (SQLite supports 3,932,160 valid options).

An experiment in which all possible permutations of option values (distinct options are known as *factors*) are tested is known as a *full factor design*. The `fac.design` function in the `DoE.base` package takes a specification of the factor levels and produces a list of all combinations that need to be run to perform a full factor design (see [experiment/design\\_fac.R](#)).

A study by Citron and Feitelson<sup>259</sup> investigated the performance impact of adding what they called a Memo-Table (essentially a cache designed to store and reuse the results of previously executed instruction sequences) to the IBM Power4 cpu architecture. The configuration options for the Memo-Table were Size (1k or 32k), Associativity (1-way or 8-way), Mapping (indexing by program counter or operand+opcode) and Replacement method (random or least recently used).

Four configuration parameters, each having two possible values, gives  $4^2 \rightarrow 16$  possible configurations. Citron and Feitelson benchmarked all 16 possibilities, enabling them to check for interactions between all factors. However, in many cases the number of interactions between factors is small and a common cost saving is to only consider interactions between pairs of factors.

Factor having just two possible values is a common case and is known as a two-factor factorial design: it is a *full two-factor design* when all combinations used and a *fractional two-factor design* when a subset is used.

The `FrF2` function, in the `FrF2` package, generates a list of the combination of factor values that need to be run to analyse  $N$  factors having a resolution of  $R$  (the ability to separate out main effects and interactions between factors; to be able to separate out main effects a resolution of 3 is required, a resolution of 4 enables detection of separate pairs of interactions). A call specifying the number of runs and number of factors produces a list of options to use for each run, along with the number of interactions that can be analysed:<sup>ii</sup>

```
> library("FrF2")
> FrF2(nfactors=4, resolution=3, alias.info=3)
  A  B  C  D
1  1  1  1  1
2 -1 -1 -1 -1
3  1 -1 -1  1
4 -1 -1  1  1
5  1 -1  1 -1
6 -1  1  1 -1
7 -1  1 -1  1
8  1  1 -1 -1
class=design, type= FrF2
```

---

<sup>ii</sup> Some functions use  $+/-$  rather than  $1/-1$ .

The price paid for running a fractional, rather than full, factorial design experiment is that it is not possible to distinguish interactions between some combinations of factors. For instance, after running the eight combinations listed above it is not possible to distinguish between an effect caused by a combination of the AB factors and one caused by the combination CD; this combination is said to be *aliased*. The complete list of aliased factors is:

```
> design.info(FrF2(nfactors=4, resolution=3, alias.info=3))$aliased
$legend
[1] "A=A" "B=B" "C=C" "D=D"

$main
[1] "A=BCD" "B=ACD" "C=ABD" "D=ABC"

$fi2
[1] "AB=CD" "AC=BD" "AD=BC"

$fi3
character(0)
```

To distinguish between an effect caused by any of these combinations, all 16 combinations of factors have to be run.

Factorial designs require the number of runs to be a power of two, so the number of different runs grows very quickly as the number of factors increases.

A Plackett and Burman design requires that the number of runs be at least one greater than the number of factors and also be a multiple of four. However, the results from experiments using these designs will only support the analysis of the main factors, i.e., any interactions between factors are ignored. Plackett and Burman designs can contain complex aliasing between the main factors and (possible) interactions between pairs of factors. Plackett and Burman designs are non-regular fractional factorial 2-level designs. The pb function in the FrF2 generates Plackett-Burman designs.

Multiple fractional factorial designs can be combined to isolate effects (i.e., remove aliasing between combinations of factors). Some signs in the original design are switched, creating what is known as a *fold over* of the original; switching the signs of all factors is known as a *full fold over*. The fold.design function generated a foldover design from an existing design.

See argument checking...

A study by Lee and Brooks<sup>826</sup> made use of three optional values per parameter; see [experiment/lee2006/lee.R](#). Randomly selecting option values is an inefficient use of resources because some option values are over/under used (see [experiment/SQLWR.R](#) from a study by Guo et al<sup>566</sup>).

### 13.3 Analysing the results

The need to compare measurement samples obtained from running experiments kick started the development of statistics. The range of possible different experimental designs (e.g., one/two/k samples, parametric/non-parametric and between/within subject) and the need for practical manual solutions produced techniques designed to handle each specific case.<sup>iii</sup>

This chapter compares measurement samples using techniques that are only practical when a computer is available to do the calculations. The two techniques are regression modeling and Monte Carlo methods (or permutation tests when the sample is small enough for it to be practical to calculate an exact answer).<sup>iv</sup>

Many data analysis techniques assume that each measurement in a sample is independent of the other measurements in the sample. There is a common kind of experiment that produces measurements likely to violate this assumption, i.e., an experiment that measures the same subject before and after the intervention.

---

<sup>iii</sup> These techniques come with requirements on the characteristics of the data, e.g., the before/after sample measurements share a common distribution, often the Normal distribution, or that samples have the same variance.

<sup>iv</sup> Other books tend to cover the manual techniques: such as the t-test, which is a special case of multiple regression using an explanatory variable indicating group membership, and the Wilcoxon-Mann-Whitney test, which is essentially proportional odds ordinal logistic regression.

The analysis of samples containing repeated measurements of the same subject is known as *within-subject*, while analysis of samples containing a single measurement of each subject is known as *between-subjects*.

Samples may be compared to check whether they are the same/different, in some sense, or by specifically testing whether one sample is greater or less than the other:

- in a *two-sided* test (also known as a *two-tailed* or *non-directional* test) the samples are checked for being the same or different, where an increase or decrease in some attribute is considered a difference. The percentage on each side, see Figure 13.4, is half the chosen p-value,
- in a *one-sided* test (also known as a *one-tailed* or *directional* test) the samples are checked for only one case, either an increase or a decrease in the measured attribute. The percentage on the one side, see Figure 13.4, is the chosen p-value,

A commonly encountered null hypothesis, when comparing two samples, is that there is no difference between them. In many practical situations a difference is expected to exist, otherwise no effort would have been invested in obtaining the data needed to perform the analysis.

Experiments are often performed because a difference in one direction is of commercial interest. However, expecting or wanting a result that shows a difference in one direction is not sufficient justification for using a one-sided statistical test.

A one-sided test should only be used when the direction is already known or when an effect in the non-predicted direction would be ignored. If an effect in a particular direction is expected, but an effect in the opposite direction would not be ignored (i.e., would be considered significant) a two-sided test should be used.

Some of the kinds of sample comparisons commonly made include:

- a level of confidence that sample values have been drawn from the same/different distribution,
- the difference,  $d_m$ , in the mean of two samples,
- the difference,  $d_v$ , in the variance of two samples,
- the correlation,  $C$ , between values paired from two samples.

**Correlated measurements** When measurements are made at specific points in time, while an experiment is running, it is possible that later measurements will be affected by earlier events that are not part of the benchmark. Perhaps running one program of a multi-program benchmark causes the system to enter a state that changes its performance characteristics.

A correlation between successive measurements, where none should exist, either needs to be removed or taken into account during analysis. The Durban Watson test can be used to check for a correlation between successive measurements within each run. The `durbinWatsonTest` function, in the `car` package implements this test. This issue is discussed elsewhere, see Figure 11.22.

### 13.3.1 Regression modeling

Regression modeling is used to analyse many of the datasets in this book. Using regression modeling to analyse experimental data might appear to be over-kill. But when a computer is available to do the calculations, it makes sense to use the most powerful analysis techniques available; why waste developer time learning to apply one of a variety of less powerful techniques (practical manual techniques are available for the different kinds of experimental data).

A study by Potanin, Damitio and Noble<sup>1107</sup> refactored the Java Development Kit collection so that it no longer made use of incoming aliases (e.g., following the owner-as-dominator or owner-as-accessor encapsulation discipline). The performance of the original and refactored versions were compared using the DaCapo benchmark,<sup>148</sup> which contains 14 separate programs, each of which iterates 30 times, with measurements made during each of the last five

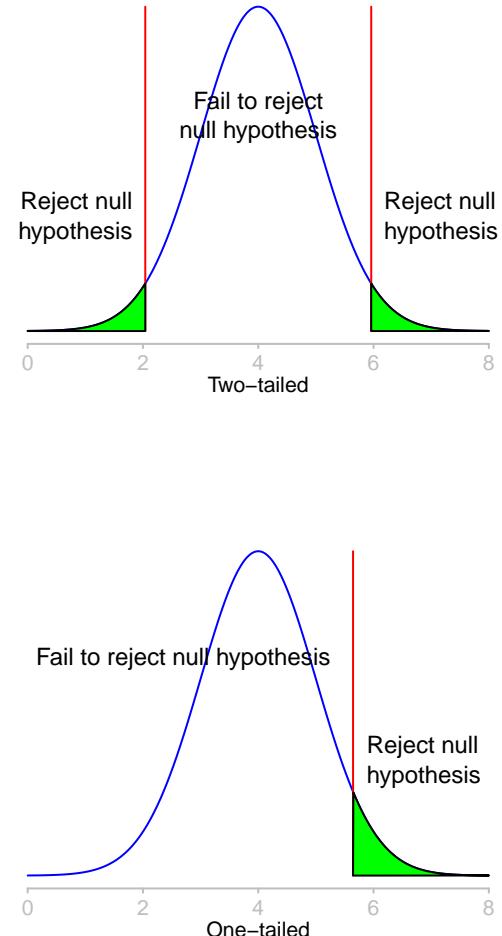


Figure 13.4: One and two-sided significance testing. `code`

iterations; this process is repeated five times, generating 25 measurements for each program for a total of 350 measurements.

Potanin et al claimed that their changes to the aliasing properties of the original code did not degrade performance. If the claim is true, the explanatory variable kind-of-refactoring will have a trivial impact on the quality of the fitted regression model. The simplest model possible is based just on the name of the program and explains 99.9% of the variance (in this case the intercept is an unnecessary degree of freedom):

```
prog_mod=glm(performance ~ progname-1, data=dacapo_bench)
```

The fitted equation essentially just contains the mean value of the runtime of each separate program, for all programs in the sample. The `summary` output is: `code`

`Call:`

```
glm(formula = performance ~ progname - 1, data = dacapo)
```

`Deviance Residuals:`

Min	1Q	Median	3Q	Max
-4174.6	-205.0	-9.0	116.6	3946.5

`Coefficients:`

	Estimate	Std. Error	t value	Pr(> t )
prognameavrora	22881.32	48.03	476.439	< 2e-16 ***
prognamebatik	2519.87	48.03	52.469	< 2e-16 ***
prognameeclipse	53660.53	48.03	1117.330	< 2e-16 ***
prognamefop	395.89	48.03	8.243	2.92e-16 ***
prognameh2	24100.39	48.03	501.823	< 2e-16 ***
prognamejython	15808.13	48.03	329.160	< 2e-16 ***
prognameluindex	708.00	48.03	14.742	< 2e-16 ***
prognamelusearch	7239.52	48.03	150.743	< 2e-16 ***
prognamepmd	4017.61	48.03	83.656	< 2e-16 ***
prognamesunflow	22788.81	48.03	474.513	< 2e-16 ***
prognametomcat	7672.11	48.03	159.750	< 2e-16 ***
prognametradebeans	27987.82	48.03	582.768	< 2e-16 ***
prognametradesoap	64888.58	48.03	1351.122	< 2e-16 ***
prognamexalan	26381.35	48.03	549.318	< 2e-16 ***
---				

`Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1`

(Dispersion parameter for gaussian family taken to be 345970)

```
Null deviance: 1.5873e+12 on 2100 degrees of freedom
Residual deviance: 7.2169e+08 on 2086 degrees of freedom
AIC: 32759
```

`Number of Fisher Scoring iterations: 2`

Apart from improving the fit for one program there is not much left to explain. Adding kind-of-refactoring as an explanatory variable (see [regression/dacapo\\_progname.R](#) for details) shows that it is not significant on its own, but some interaction exists between a few programs (primarily sunflow) and some refactorings. The slightly more complicated model:

```
prog_refact_mod=glm(performance ~ progname+progname:refact_kind,
                     data=dacapo_bench)
```

explains 99.92% of the variance. There are 12 program/refactoring interactions with p-values less than 0.05 (out of 84 possible interactions), with most of these changing the estimated mean performance by around 1% and one making 8% difference (sunflow, see [\[regression/-dacapo\\_progname\\_refact.R\]](#)).

Building a regression model has enabled us to confirm that apart from a few, small, interactions the various refactorings of the JDK did not change the DaCapo benchmark performance.

### 13.3.2 Factorial designs

A variety of techniques have been created to help visualise the results obtained from experiments using a factorial design. The analysis is the same for full and partial fractional designs,

the only difference is the number of interactions between factors that will be available for analysis.

The study by Citron and Feitelson<sup>259</sup> discussed earlier used the SPEC CPU 2000 benchmark, and the results measured were integer floating-point performance, power consumption, processor timing and die area of the chip.

For simplicity consider three of the factors (ignoring, for the time being, the replacement method), the results can be visualised as a cube with each of the eight vertices representing one combination of factor values. In Figure 13.5 the values at each vertex are the SPEC benchmark cint performance figures.

Imagine taking two opposite faces of the above cube, say the two for size on the left and right going into the page, and finding the mean cint value for both faces, the difference between these two values is known as the *main effect* for size; this calculation can be repeated to find the main effect for the other factors.

A design plot is a visualisation of these main effects, with a central horizontal line showing the overall mean value of the response variable. Figure 13.6 was created using the `plot.design` function and shows the impact that each factor can have on the value of cint, offset from the mean value.<sup>v</sup>

For an example involving more changeable parameters see [experiment/lee2006/lee.R](#).

At a finer level of granularity, an interaction plot shows the interaction between pairs of factors. The upper plot in Figure 13.7 shows how the mean value of cint varies as size is changed, for a given value of associativity (see legend); the lower plot shows the variation for given values of mapping.

For an equation based analysis, a regression model could be used to investigate the interactions between factors. For instance, the following code specifies interactions between all variable pairs (see [experiment/MemoPower03.R](#) for details):

```
Memo_glm=glm(cint ~ (size+associativity+mapping)^2, data=Memo)
```

A study by Pallister, Hollis and Bennett<sup>1050</sup> investigated the power consumed by various embedded programs when compiled with gcc using various command line parameters... Thirty six [experiment/pallister/gcc-power.R](#)...

Plackett and Burman designs do not contain enough information to build a regression model and a bespoke method has to be used. The `DanielPlot` function in the `FrF2` package can be used to perform the calculation and plot the results; the x-axis shows effect size, the y-axis contains diagnostic information. If the data is simply the result of random variation (i.e., changing factor values has no effect), differences between pairs of factor averages would have a (roughly) normal distribution; plotting values from a normal distribution using a normal probability scale produces a straight line. If many of the points in a `DanielPlot` appear to form a straight line, then the corresponding factors are likely to have had little effect on the results; those factors well off the line are of interest.

A study by Debnath, Mokbel and Lilja<sup>328</sup> investigated the impact of seven system configuration settings on PostgreSQL performance on the TPC-H benchmark. High and low values were chosen for the configuration values and a Plackett and Burman design with full fold-over was used. Figure 13.8 shows the half-normal plot from the 16 runs; factors P4 and P7 do not fall on the line through the other factors and also exhibit the largest effect.

### 13.3.3 Comparing sample means

Comparing two samples to check for any difference in their mean values is probably the most common statistical test of experimental data.

In the past the appropriate statistical technique to use has depended on the design of the experiment performed (e.g., within-subjects or between-subjects), the kind of question being asked and the characteristics of the data. Now that computers are available to do the calculation, the bootstrap has become the hammer used to answer sample comparison questions.

The bootstrap procedure often starts by assuming there is no difference, in some characteristic, between samples; it then calculates the likelihood of two samples having characteristic

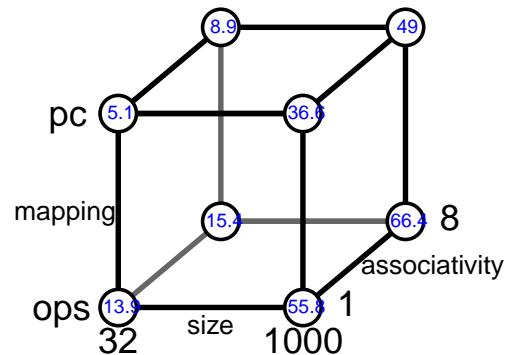


Figure 13.5: A cube plot of three configuration factors and corresponding benchmark results (blue) from Memory table experiment. Data from Citron et al.<sup>259</sup> [code](#)

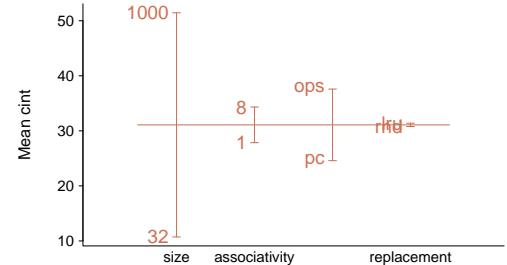


Figure 13.6: Design plot showing the impact of each configuration factor on the performance of Memo table on benchmark performance. Data from Citron et al.<sup>259</sup> [code](#)

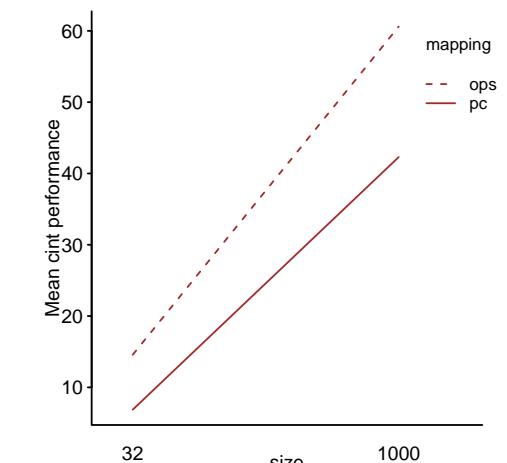
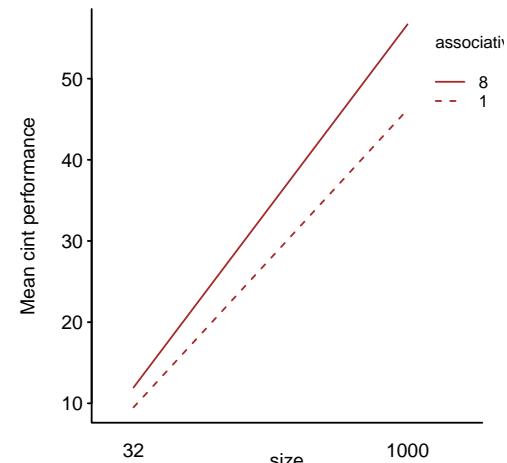


Figure 13.7: Interaction plot showing how cint changes with size for given values of associativity and mapping. Data from Citron et al.<sup>259</sup> [code](#)

<sup>v</sup> `plot.design` makes some unexpected display decisions when the explanatory variables are not factors.

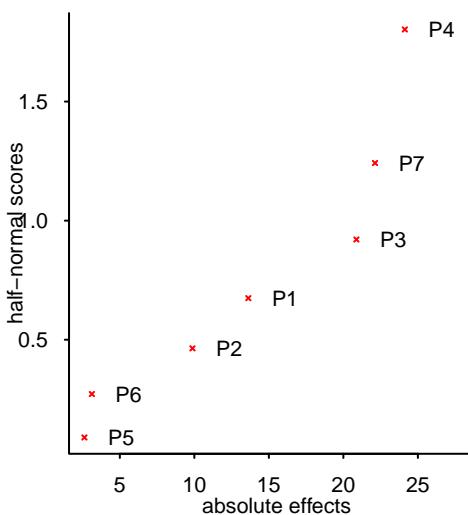


Figure 13.8: Half-normal plot of data from a Plackett and Burman design experiment. Data from Debnath et al.<sup>328</sup> code

that they are measured to have. The assumption of no difference requires that the items in both samples be *exchangeable*. Deciding which items, if any, in a sample are exchangeable is a crucial aspect of using the bootstrap to answer questions about samples.

The nVidia GTX 970 is a very popular graphics card and many variations on the reference design have been produced (during August 2016 there were 51 variants included in the 64,392 results for this card in the [UserBenchmark.com](#) database). Figure 13.9 shows the number of Reflection benchmark results reported for GTX 970 cards from three third-party manufacturers.

The mean score of these Asus, MSI and Gigabyte cards are 176.2, 179 and 186.8 respectively. Are these differences most likely caused by random variation or by some real difference?

The following bootstrap technique provides a way of answering this question.

Assume there is no difference in the mean performance of, say, MSI and Gigabyte on the Reflection benchmark. In this case the benchmark results (255 from MSI and 73 from Gigabyte) can be merged to form a sample of 328 results. Using this combined empirical sample perform the following:

- randomly select, with replacement, 328 items from the empirical sample,
- divide this new sample into two subsamples, one containing 255 items and the other 73 items,
- find the mean of the two subsamples, subtract the two mean values and record the result,
- repeat this process  $R$  times,
- count how many bootstrapped differences in the mean are greater than the differences in the means of the two cards; no assumption is made about the direction of the difference, i.e., this is a two-sided test.

The following code uses the `boot` function, from the `boot` package, to implement the above algorithm, with the user provided function (`mean_diff` in this case) that is called for each randomly generated sample (see [group-compare/UserBenchmark\\_compare.R](#)):

```
library("boot")

mean_diff=function(res, indices)
{
  t=res[indices]
  return(mean(t[1:num_MSI])-mean(t[(num_MSI+1):total_reps]))
}

MSI_refl=MSI_1462_3160$Reflection
Giga_refl=Gigabyte_1458_367A$Reflection

num_MSI=length(MSI_refl)
num_Giga=length(Giga_refl)
total_reps=num_MSI+num_Giga

GTX_boot=boot(c(MSI_refl, Giga_refl), mean_diff, R = 4999)

refl_mean_diff=mean(MSI_refl)-mean(Giga_refl)
# Two-sided test
length(GTX_boot$t[abs(GTX_boot$t) >= abs(refl_mean_diff)]) # E
```

The argument  $R$  specifies the number of resamples and `boot` returns the result of calling `mean_diff` for each of these samples.

The likelihood of encountering a difference in mean values as large as that seen in the MSI and Gigabyte performance (i.e., the p-value) is given by the following equation:

$$\frac{E + 1}{R + 1}$$

where  $E$  is the number of cases where the bootstrap sample had a larger mean difference. The result varies around:  $\frac{34+1}{4999+1} \rightarrow 0.007$  (the MSI/Asus comparison the value is:  $\frac{840+1}{4999+1} \rightarrow 0.17$ ).

If there were no difference in performance, a difference in mean value as large as that seen for MSI/Gigabyte is likely to occur 0.7% of the time; we might reasonably claim that this percentage is so small that there is likely to be a real difference in performance. A mean difference at least as large as the MSI/Asus mean difference is likely to occur 17% of the time if there was no real difference in performance; a large enough percentage to infer that there is unlikely to be any difference in performance.

This test answers the question of whether the difference in mean values is likely to be a chance effect.

If a difference is now thought likely to exist, the next question is the likely size of the difference and the confidence intervals on this difference.

A bootstrap procedure can be used to answer these questions.

Once the two samples are considered to be different, their contents can only be treated as exchangeable within each sample, not between the two samples. The two subsample now have to be generated from their respective empirical samples. The following code implements the functionality (see [group-compare/UserBenchmark\\_mdiff.R](#)):

```
library("boot")

mean_diff=function(res, indices)
{
  t=res[indices, ]
  return(mean(t$refl[t$vendor == "Gigabyte"])- mean(t$refl[t$vendor == "MSI"]))
}

# Create dataframe identifying vendor used for each measurement.
MSI_refl=data.frame(vendor="MSI", refl=MSI_1462_3160$Reflection)
Giga_refl=data.frame(vendor="Gigabyte", refl=Gigabyte_1458_367A$Reflection)

MSI_Giga=rbind(MSI_refl, Giga_refl)

# Pass combined dataframe and specify identifying column
GTX_boot=boot(MSI_Giga, mean_diff, R = 4999, strata=MSI_Giga$vendor)
```

The `boot.ci` function calculates confidence intervals from the value returned by `boot`: [code](#)

```
> boot.ci(GTX_boot)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 4999 bootstrap replicates

CALL :
boot.ci(boot.out = GTX_boot)

Intervals :
Level      Normal          Basic
95%   ( 4.537, 11.166 )  ( 4.488, 11.069 )

Level      Percentile        BCa
95%   ( 4.560, 11.141 )  ( 4.982, 11.775 )
Calculations and Intervals on Original Scale
> mean(GTX_boot$t)
[1] 7.777926
> sd(GTX_boot$t)
[1] 1.6911
```

Deciding if and when items in a sample are exchangeable can be non-trivial and require an understanding of the problem domain.

A study by Gandomani, Wei and Binhamid<sup>473</sup> investigated the accuracy of software cost estimates made using both expert judgement and Planning Poker on 15 projects in one company and both expert judgement and Wideband Delphi in 17 projects in another company; Table 13.1 shows a subset.

Is there a difference in the estimates made using expert judgement and either of the other two techniques?

Project	Expert judgement	Planning Poker	Difference
P1	41	40	1
P4	60	56	4
P7	33	45	-12
P12	18	20	-2

Table 13.1: Effort estimates made using expert judgement and Planning Poker for several projects. Data from Gandomani et al.<sup>473</sup>

Each estimate is specific to one project and it makes no sense to include estimates from other projects in the random selection process; estimates from different projects are not exchangeable. Possible ways of handing this include:

- treating each project as being exchangeable; the resampling could occur at the project level with both estimates for each selected project being used.
- randomly selecting from the set of estimates for each project. But there are only two estimates...

The following code randomly samples estimates for each project only within the sample of estimates for that project (see [group-compare/16.R](#)):

```
mean_diff=function()
{
  s_ind=rnorm(len_est_2) # random numbers centered on zero
  # Randomly assign estimates to each group
  expert=c(est_2$expert[s_ind < 0], est_2$planning.poker[s_ind >= 0])
  # Sampling with replacement, so two sets of random numbers needed
  s_ind=rnorm(len_est_2) # random numbers centered on zero
  poker=c(est_2$expert[s_ind < 0], est_2$planning.poker[s_ind >= 0])
  # The code for sampling without replacement
  # poker=c(est_2$expert[s_ind >= 0], est_2$planning.poker[s_ind < 0])
  return(mean(expert)-mean(poker))
}

est_mean_diff=abs(mean(est_2$expert)-mean(est_2$planning.poker))
len_est_2=nrow(est_2)

t=replicate(4999, mean_diff()) # Implement our own bootstrap
length(t[abs(t) > est_mean_diff]) # How many this extreme?
```

The p-value for Expert vs. Planning Poker is 0.02 (see [group-compare/16.R](#))....

A study by Jørgensen and Carelius<sup>713</sup> asked companies to bid on a software development project.<sup>vi</sup> In the first round of bidding 17 companies were given a one-page description of user needs and asked to supply a non-binding bid; in the second round the original 17 companies plus an additional 18 companies (who had not participated in the first round) were given an 11-page specification (developed based on feedback from the first round) and asked to submit firm-price bids.

What difference, if any, did participating in the first round make to the second bids submitted by the initial 17 companies (call them the A companies) and how did these bids compare to those submitted by the second sample of 18 companies bidding for the first time (call them the B companies)?

Figure 13.10 shows density plots of the submitted bids. The mean values were: mean of kr183,051<sup>vii</sup> for initial bid from A companies, mean of kr277,730 for final bid from A companies and mean of kr166,131 for only bid from B companies.

Are the items in each sample (the companies asked to submit a bid) exchangeable? Small companies have lower operating costs than large companies; it is unrealistic to consider bids from small/large companies to be exchangeable. The size of companies involved in bidding

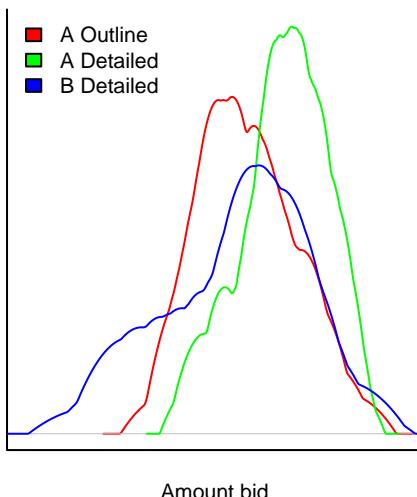


Figure 13.10: Density plots of project bids submitted by companies before/after seeing a requirements document. Data from Jørgensen et al.<sup>713</sup> [code](#)

<sup>vi</sup> Four of the companies that submitted a bid were selected to independently implement the project.

<sup>vii</sup> The exchange rate is approximately 10 Norwegian Krone to one Euro.

were classified as small (five or fewer developers), medium (between 6 and 49 developers) and large (50 or more developers).

The call to `boot` now has to include information on how the data is stratified (i.e., split into different levels). The `strata` is used to pass a vector of integer values specifying the strata membership of the values in the first argument. Everything else stays the same, with `boot` treating members of each strata as exchangeable when generating new samples (see [group-compare/compare-bid.R](#)):

```
bid_boot=boot(comp_bid$Bid, mean_diff, R = 4999,
              strata=as.factor(comp_bid$CompSize))
```

The p-value, for the hypothesis that the mean values are the same, is:  $\frac{52+1}{4999+1} \rightarrow 0.01$ , i.e., a difference this large is surprising.

Jørgensen and Carelius proposed the hypothesis that the main factor controlling the size of the bids was the information contained in the project specification. I think this is rather idealistic and more practical considerations are discussed in Section [5.3](#).

Within subjects experiments using bootstrap... Given two samples,  $S_1$  where X was not performed and  $S_2$  where X was performed, the mean performance time of the samples, for instance, is likely to be different. How surprised should we be at this difference?

**Permutation tests** Sometimes the two samples are small enough that it is practical to generate and test all possible item permutations.

A study by Grant and Sackman<sup>538</sup> measured the time taken for subjects to write a program using either an online or offline computer interface (this experiment was run in the mainframe era of the 1960s). Given 12 subjects split into two groups of six, how likely is the difference in mean time between the online/offline cases?

This question is about the population of people that took part in the experiment, not a wider population. For this population there are `choose(12, 6) == 924` possible subject combinations. The following is an excerpt of an implementation of a two-sided test (see [group-compare/GS-perm-diff.R](#)):

```
subj_time=c(online$time, offline$time)
subj_mean_diff=mean(online$time)-mean(offline$time)

# Exact permutation test
subj_nums =seq(1:total_subj)
# Generate all possible subject combinations
subj_perms=combn(subj_nums, subj_online)

mean_diff = function(x)
{
  # Difference in mean of one combination of subjects
  mean(subj_time[x]) - mean(subj_time[!(subj_nums %in% x)])
}

# Indexing by column iterates through every permutation
perm_res=apply(subj_perms, 2, mean_diff)

# p-value of two-sided test
sum(abs(perm_res) >= abs(subj_mean_diff)) / length(perm_res)
```

For the Algebra program, 272 of the possible groups, of subject combinations, had a difference in mean time greater than, or equal, to that of the empirical sample. Because all possibilities have been calculated, the p-value is exact:  $\frac{272}{924} \rightarrow 0.2934....$

The `coin` package provides this kind of exact calculation for many of the traditional group comparison tests, e.g., the `wilcoxonsign_test` function is permutation test equivalent of the `wilcox.test` function in the base library.

Clustering by mean value using Scott-Knott test... `ScottKnott`

The bootstrap techniques applied to question about differences in the mean of two samples can be generalised to a wide variety of comparison tests. A new comparison test can be implemented by replacing the `mean_diff` function used above (the requirement of exchangeability is integral to the process).

When comparing different systems, benchmark performance may be normalised to produce a relative performance ranking. The geometric mean has to be used when comparing normalized values, otherwise the results can be inconsistent.

Table 13.2 shows the results of five benchmark programs from three systems (i.e., columns R, M and Z); two other sets of columns list normalised values, with different processors used as the reference. The bottom two rows list the arithmetic and geometric mean of the columns. A ranking based on the arithmetic mean depends on the processor used as the base for normalization, while the geometric mean produces a consistent ranking.

	<b>R</b>	<b>M</b>	<b>Z</b>	<b>R/M</b>	<b>M/M</b>	<b>Z/M</b>	<b>R/R</b>	<b>M/R</b>	<b>Z/R</b>
E	417.00	244.00	134.00	1.71	1.00	0.55	1.00	0.59	0.32
F	83.00	70.00	70.00	1.19	1.00	1.00	1.00	0.84	0.84
H	66.00	153.00	135.00	0.43	1.00	0.88	1.00	2.32	2.05
I	39449.00	33527.00	66000.00	1.18	1.00	1.97	1.00	0.85	1.67
K	772.00	368.00	369.00	2.10	1.00	1.00	1.00	0.48	0.48
Arithmetic	8157.40	6872.40	13341.60	1.32	1.00	1.08	1.00	1.01	1.07
Geometric	586.79	503.13	498.68	1.17	1.00	0.99	1.00	0.86	0.85

Table 13.2: Benchmark results from three processors, with normalisation using two different processors. Data from Fleming et al.<sup>446</sup> [code](#)

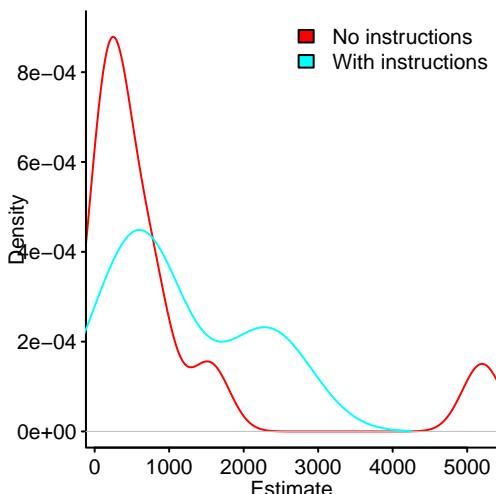


Figure 13.11: Density plot of task implementation estimates: with no instructions (red) and with instruction on what to do (blue). Data from Jørgensen el al.<sup>717</sup> [code](#)

### 13.3.4 Comparing standard deviation

A study by Jørgensen and Moløkken<sup>717</sup> asked 19 professional developers to estimate the effort required to implement a task, along with an uncertainty estimate (i.e., minimum and maximum about the most likely value). Nine of the developers were explicitly instructed to compare the current task with similar projects they had worked on (they were also given a table that asked them to assess similarity within various percentage bands).

The visual appearance of the density plots in Figure 13.11 suggests that there is a difference in the standard deviation of the estimates in the two samples. A bootstrap test of the difference in the standard deviations of the two samples can be implemented by replacing the `mean_diff` used in the previous section by the function `sd_diff` below (see [group-compare/simula\\_04sd.R](#)):

```
sd_diff=function(est, indices)
{
  t=est[indices]
  return(sd(t[1:num_A_est])-sd(t[(num_A_est+1):total_est]))
}
```

The p-value, for the hypothesis that the standard deviations are the same, is:  $\frac{2170+1}{4999+1} \rightarrow 0.43$ , i.e., the difference is not that surprising.

The `ansari.test` function of the `coin` package<sup>viii</sup> implements a permutation test of the *Ansari-Bradley Test* (a two-sample test for a difference in variance); see [group-compare/simula\\_04\\_v](#)

Median Absolute Deviation `mad` for robust estimation of variance... Winsorized variance...

### 13.3.5 Correlation

Correlation is a measure of the closeness of linear association between variables, e.g., if one variable always increases/decreases when another variable increases/decreases. The range of correlation values is -1 (the variables change together, but in opposite directions) to 1 (the variables always change together), with zero denoting no correlation.

Correlation is related to regression, except that: it treats all variables equally (i.e., there are no response or explanatory variables), the correlation value is dimensionless and correlation is a linear relationship (e.g., in the  $y = x^2$  relationship  $y$  can be predicted  $x$ , but there is zero correlation between them).

<sup>viii</sup> The `ansari.test` function is included in R's base system.

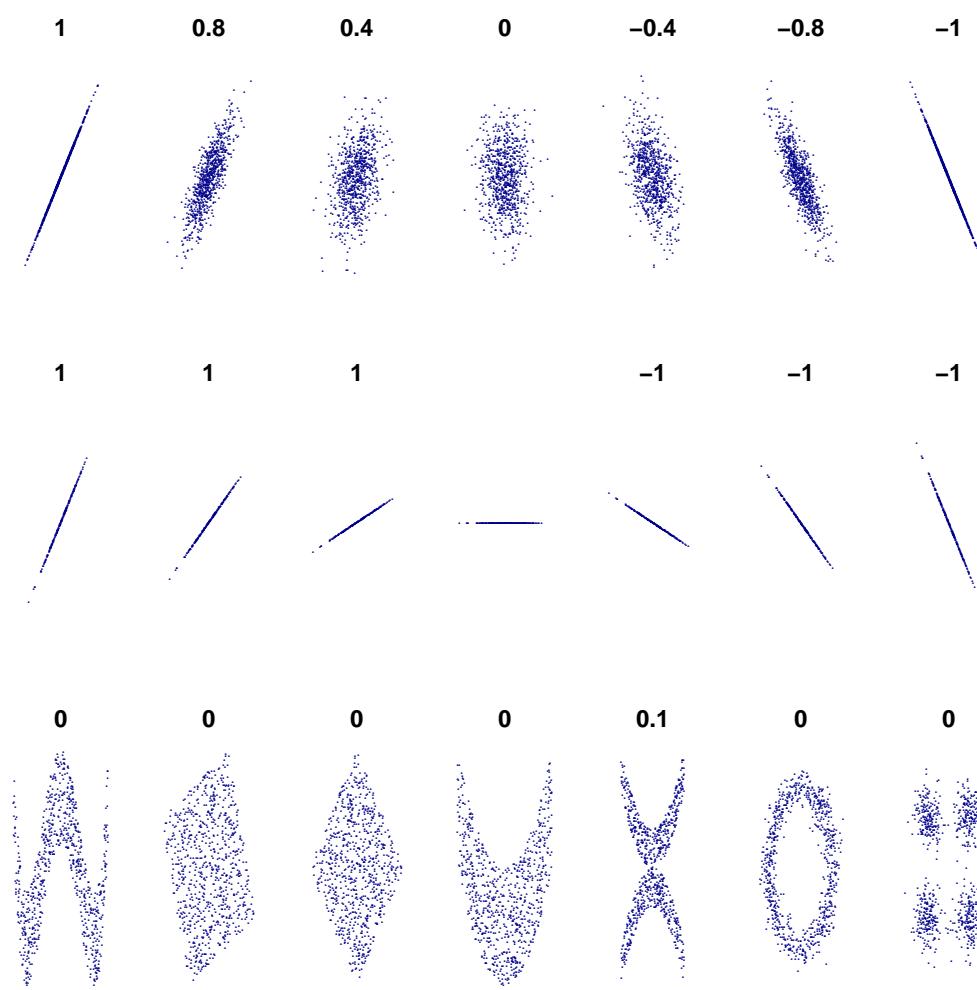


Figure 13.12: Examples of correlation between samples of two value pairs, plotted on x- and y-axis. [code](#)

Three commonly encountered metrics for correlation are:

- *Pearson product-moment correlation coefficient*, (also known as Pearson's R or Pearson's r), which applies to continuous variables,
- Spearman's rho,  $\rho$  (a lowercase Greek letter), is identical to Pearson's coefficient except the correlation is calculated from the ranked values, i.e., the sorted order (which makes it immune to extreme values),
- Kendall's tau,  $\tau$  (a lowercase Greek letter), is like Spearman's rho in that it is based on ranked values, but the calculation is based on the number of items sharing the same rank (i.e., relative difference in rank plays is not included in the calculation; Spearman's rho does include relative differences).

Because they involve ranking items both Spearman's rho and Kendall's tau have problems handling samples containing items having the same value...

The `cor.test` function, included in the base system, supports all three coefficients and provides confidence interval.

### 13.3.6 Dichotomous variables

When the result of a measurement can only have one of two values, the standard techniques for calculating correlation, which require that most if not all values be unique, cannot be used. It is possible to recast the problem in terms of probabilities, which means that the approach taken for every problem could be different.

The following is an example of one approach to a particular binary problem.

If we are interested in being able to access files with very high reliability, then hosting the files on two or more websites would mean that if one site cannot be accessed, the file could

be obtained from another site. The naive analysis suggests that if the average reliability of the websites is 95% then the reliability of two paired sites would be 99.75%; however this assumes that the unavailability of each website is independent of its paired site.

A study by Bakkaloglu, Wylie, Wang and Ganger<sup>89</sup> had a client program read a file from over 120 websites every 10-minutes, between September 2001 and April 2002. They recorded whether the file was successfully accessed or not.

Most of the websites are available most of the time, and this makes it harder to detect if other patterns are present. Bakkaloglu et al proposed various techniques for calculating correlated failures, based on the probability that site  $X$  is unavailable when site  $Y$  is unavailable, i.e.,  $P(X\text{unavailable}|Y\text{unavailable})$ . The following example takes the mean value over all pairs of sites:

$$\begin{aligned} &= \text{mean}(P(X\text{unavailable}|Y\text{unavailable})) \\ &= \text{mean}\left(\frac{P(X\&Y\text{unavailable})}{P(Y\text{unavailable})}\right) \end{aligned}$$

The following calculates the average unavailability probability for one site paired with every other site (see [probability/reliability/web-avail.R](#)):

```
given=web_down[ , ind]
others=web_down[ , -ind]

both_down=(others & given)

av_prob=mean(colSums(both_down)/sum(given))
```

Averaged over all pairs of sites the probability of one site being unavailable when its pair is also unavailable is 0.3 (at the 10-minute measurement point). Given that all accessed originated from the same client it is not surprising that this probability is much higher than the average probability of one site being unavailable (0.1); all accesses start off going through the same internet infrastructure and problems with this infrastructure will affect access to all sites.

### 13.3.7 Contingency tables

Count data with categorical explanatory variables has a natural visual representation as a table of numbers; these tables are known as *contingency tables*. Table 13.3 shows a 2-dimensional table, with each entry containing a count of the items in the sample having both of the listed row and column attributes.

It is surprisingly common to encounter situations where lots of data has been reduced to a contingency table, i.e., potentially useful information has been thrown away. The only reason for reducing lots of data to the form of a contingency table is to hide information from readers. Analysis the whole sample is likely to reveal more about it than an analysis of a simplified form, i.e., analysis of a contingency table.

Sometimes the available measurements are only sufficient to build a contingency table.

A study by Nightingale, Douceur and Orgovan<sup>1009</sup> investigated the characteristics of hardware failures over a very large number of consumer PCs. Table 13.3 shows a contingency table containing the number of system crashes believed to have been caused by hardware problems involving the system DRAM or CPU.

	<b>DRAM failure</b>	<b>no DRAM failure</b>
<b>CPU failure</b>	5	2,091
<b>no CPU failure</b>	250	971,191

Table 13.3: Number of system crashes of consumer PCs traced to CPU or DRAM failures. Data from Nightingale et al.<sup>1009</sup>

The traditional, manual friendly, technique for analyzing this kind of data is the chi-squared test ( $\chi$  is the Greek letter), which provides a yes/no answer.<sup>ix</sup>

<sup>ix</sup> The `chisq.test` function is part of the base system if your readership demands a chi-squared test and the `chisq_test` function in the `coin` package can be used to bootstrap confidence intervals.

A regression model can be fitted to this data (even though there is not a lot of it) and can extract a more information than the manual method. [code](#)

Call:

```
glm(formula = failures ~ CPU * DRAM, family = poisson, data = PC_crash)
```

Deviance Residuals:

```
[1] 0 0 0 0
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	13.786278	0.001015	13586.243	< 2e-16 ***
CPUTRUE	-6.140881	0.021892	-280.505	< 2e-16 ***
DRAMTRUE	-8.264818	0.063254	-130.661	< 2e-16 ***
CPUTRUE:DRAMTRUE	2.228858	0.452194	4.929	8.27e-07 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2.6646e+06 on 3 degrees of freedom

Residual deviance: -7.7825e-11 on 0 degrees of freedom

AIC: 43.948

Number of Fisher Scoring iterations: 3

The information in the fitted model that is not immediately obvious from the numbers in the table is that the crash rate is higher when both the CPU and DRAM fail.

The regression approach makes it trivial to handle more rows and columns, as well as non-straight line fits. As the number of values in the table increases it becomes more difficult to visually extract any patterns that may be present; a heatmap is one way of highlighting details, see Figure 13.13.

There are a wide variety of techniques for comparing multiple contingency tables. Different pair comparison algorithms can give very different results. [1337](#)

### 13.3.8 ANOVA

Readers are likely to encounter the acronym ANOVA (*Analysis of variance*), an analysis technique which developed independently of linear regression and having its own specialized terminology. This technique was designed for manual implementation.

Functionally ANOVA and least squares are both special cases of the general linear model (ANOVA is a special case of multiple linear regression with orthogonal, categorical predictors; ANCOVA adds covariates to mix). A one-way analysis of variance can be thought of as a regression model having a single categorical predictor that has at least two (usually more) categories.

Treating the various kinds of ANOVA models as special cases of the family of regression models makes it possible to use the more flexible options available in regression modeling (e.g., easier handling of unequal group sizes, adjusting for covariates and methods for checking models).

The `anova` function generates ANOVA style output when passed a model built using `glm` and some other regression model building functions. The `Anova` function in the `car` package...

One-way ANOVA focuses on testing for differences among a group of means; it evaluates the hypothesis that  $\alpha_i = 0$  in the following equation:

$$Y_i = \mu + \alpha_i + \varepsilon_i$$

where  $\mu$  is the group mean,  $\alpha_i$  is the effect of the response variable on the  $i$ 'th group and  $\varepsilon_i$  is the corresponding error.

The bootstrap is a more flexible and powerful technique.

`Anova` can also be used to compare linear models...

The `glht` function provides a comprehensive set of methods for multiple mean comparisons, works for GLM...

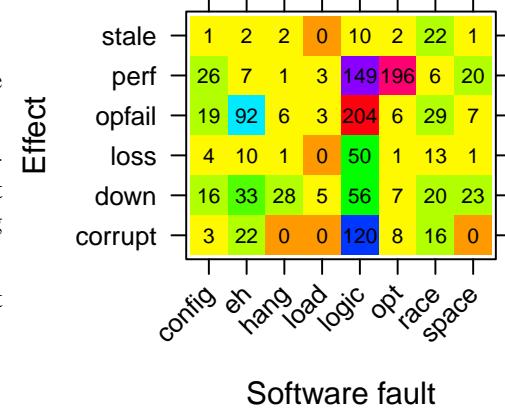


Figure 13.13: Number of software faults having a given consequence, based on an analysis of faults in Cassandra. Data from Gunawi et al. [562](#) [code](#)

## 13.4 Benchmarking

Benchmarking is the process of running an experiment to obtain information about some aspect of hardware and/or software performance. The performance measure of interest may be as simple as checking that the system completes a specified task within a given amount of time or may involve complex measurements of resource usage.

Common reasons for benchmarking in software engineering include comparing before/after performance and obtaining numbers to put in a report. For a more general audience benchmarking information is often used as input to a selection process and as such often has a marketing orientation. Accurate measurements are not always necessary, showing that a system is good enough may be good enough.

The time taken for operations such as add and multiply was used to compare the performance of early computers.<sup>147,928,1439,1440</sup> Studies by Knight<sup>773,774</sup> calculated performance based on a weighted average of instruction times, based on the kinds of instructions executed by commercial and scientific programs. Based on a study of over 300 computers available between 1944 and 1967 the rental cost for performing an operation decreased with increasing computer performance; a power law with an exponent between two and three (see Figure 13.14 and also [benchmark/EvolvingCompPerf\\_1963-1967.R](#)).

On modern computing platforms, obtaining accurate benchmark data for many kinds of question is likely to be economically infeasible.<sup>x</sup> A consequence of the continual reduction in the size of components within microprocessors, see Figure 13.15 and Figure 11.49, is that individual components are now so small that variations in the fabrication process (e.g., differences in the number of atoms added or removed during the fabrication process) can noticeably change their geometry, leading to large variations in the runtime electrical characteristics of supposedly identical devices.<sup>134</sup>

Manufacturers offer computing systems having a range of different performance characteristics; the lower plot in Figure 13.16 shows all published results for the integer SPEC2006 benchmark. While hardware performance has now improved to the point where for many uses it appears to be good enough, it is still possible to buy hardware that is under-powered for the job it is expected to perform (the upper plot in Figure 13.16 shows the orders of magnitude improvements in cost and performance of sorting over 15 years).

Benchmark results published for general consumption should be treated with extreme caution, they are often little more than marketing information. The author(s) may have reasons for wanting to create a favourable impression for one system in preference to others or may just have done a sloppy job (because of inexperience, incompetence or lack of resources). One class action suite alleged that:<sup>490</sup> “Intel used its enormous resources and influence in the computing industry to, in Intel’s own words, “falsely improve” the Pentium 4’s performance scores. It secretly wrote benchmark tests that would give the Pentium 4 higher scores, then released and marketed these “new” benchmarks to performance reviewers as “independent third-party” benchmarks. It paid software companies to make covert programming changes to inflate the Pentium 4’s performance scores and even disabled features on the Pentium III so that the Pentium 4’s scores would look better by comparison.”<sup>xi</sup>

In some consumer goods markets, product benchmark results receive a lot of publicity, with potential customers thought to be influenced by the results achieved by similar products. A study by Shimpi and Klug<sup>36</sup> of Android benchmarks found that some mobile phone vendors detected when a particular benchmark was being run and raised the devices thermal limits (allowing the system clock rate to run faster for longer; a 4.4% performance improvement was measured).

Researchers are happy to complain about poor benchmarking practices, but are not always willing to name names.<sup>1387</sup>

In published benchmark results the Devil is in the detail, or more often in the lack of detail, as illustrated by the following:

- Bailey<sup>84</sup> lists twelve ways in which parallel supercomputer benchmarks have been written in a way likely to mislead readers, including: quoting 32-bit, not 64-bit results, quoting

<sup>x</sup> Obtaining accurate benchmark results has always been an expensive and time-consuming process, but at least it used to be possible to count on devices sharing the same part number having the same performance characteristics.

<sup>xi</sup> The class action was settled<sup>1318</sup> with Intel agreeing to pay \$15 to Pentium 4 purchasers, \$4 million to a non-profit entity and an amount not to exceed \$16.45 million to the lawyers who brought the suit.

figures for the inner kernel of the computation as if they applied to the complete application and comparing sequential code against parallelized code.

- Citron<sup>258</sup> analysed the ways in which many research papers using the SPEC CPU2000 suite have produced misleading results by only using a subset of the benchmark programs (only 23 out of 115 papers surveyed used the whole suite). In one case a reported speed up of 1.42 is reduced to 1.16 when the whole suite is included in the analysis (reduction from 1.43 to 1.13 in another and from 1.76 to 1.15 in a third).

The primary purpose of this section is to highlight the many sources of variability present in modern computing systems. The available evidence suggests that large variations in benchmark results are now the norm. Large variations in measured performance do not prevent accurate results being obtained, they increase the time and money needed (i.e., it is simply a case of making enough measurements). Advice on how to perform benchmarks is available elsewhere.<sup>417, 565</sup>

### 13.4.1 Following the herd

When choosing a benchmark, there is a lot to be said for doing what everybody else does, advantages include:

- can significantly reduce the cost and time needed to obtain benchmark data,
- an established benchmark is likely to be usable out-of-the-box. It takes time for a benchmark to become established; an analysis<sup>1071</sup> of one Java source code corpora was only able to build 86 of the 106 Java systems in the corpus (and 56 of these had to be patched to get them to build),
- it is an easier sell to the result to audiences when the benchmark used is known to them.

The problem with following the herd is that there may be fitness-for-purpose issues associated with using the benchmark, i.e., herd behavior is adapted to environments that may be substantially different from the environment in which the system will operate. For instance, the SPEC benchmark is often used for comparing compiler performance, while SPEC is designed for benchmarking processor performance not compiler performance.

Commonly used benchmarks suffer from vendors tuning their products to perform well on the known characteristics of the benchmark. The SPEC benchmark has been used over many years for compiler benchmarking and compiler vendors often use it in-house for performance regression testing.

### 13.4.2 Variability in today's computing systems

In the good old days, computer performance tended to be relatively consistent across identical, but physically different, components, i.e., the same model of cpu or memory chip. Also, software often had relatively few options that could significantly alter its performance characteristics.

Components in modern hardware may be fabricated using handfuls of atoms and process variations of an atom or two here and there can produce devices that are apparently identical, but have surprisingly different performance characteristics.<sup>960</sup> Further reductions in the number of atoms used to fabricate devices will lead to greater variations in the final product. Today's consumer of benchmark results has to chose between:

- accepting a wide margin of error,
- requiring that a benchmark be executed very many times to ensure there are enough measurements to obtain the desired statistical confidence interval for the results; this approach also involves checking that a wide range of, possible unknown, factors are controlled for by those running the benchmark.

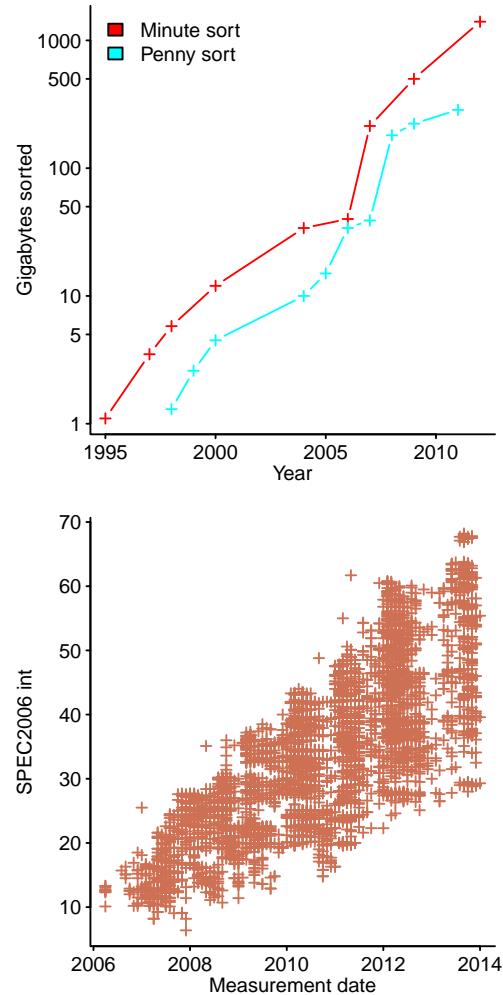


Figure 13.16: Maximum number of records sorted in 1 minute and using 1 penny's worth of system time (upper). SPEC2006 integer benchmark results (lower). Data from Gray et al<sup>542</sup> and SPEC.<sup>1282</sup> code

Intrinsic variability in system performance has implications for companies that regularly monitor the performance of their product during ongoing development. For instance, Mozilla regularly measures the performance of the latest checked-in version of source code of Firefox, if an update results in a decrease in performance that exceeds a predefined limit, the update is rolled back. Successful implementation of such a policy requires that the impact of external factors on performance are carefully controlled.

Performance variation has to be addressed from a system wide perspective,<sup>xii</sup> hardware-/software interaction can have a significant performance impact and there are often multiple, independent, sources of variation.

At the systems level differences in component characteristics (e.g., differences in system clock frequency drift in multiprocessor systems<sup>653</sup>) can interact to produce emergent effects.

DVFS (Dynamic Voltage and Frequency Scaling) provides an example of how the complexities of system component interactions make it difficult to reliably predict performance. As its name suggests DVFS allows processor voltage and frequency to be changed while programs are running and an analysis of system power consumption<sup>1479</sup> concludes that total power consumed, executing a program from start to finish, is minimised by running the processor as fast as possible (assuming there is no waiting for user input).

A study by Götz, Ilsche, Cardoso, Josef Spillner, Aßmann, Nagel and Schill<sup>528</sup> investigated how the total system power consumed by implementations of various algorithms varied with cpu clock frequency, with the intent of finding the frequency which minimised power consumption.

Figure 13.17 shows that total power consumption does not decline with frequency, there is a frequency below the maximum that minimises power consumed.<sup>1413</sup> The power minimisation frequency depends on the implementation of the sorting algorithm and the difference between minimum and maximum depends on the number of items being sorted. Predicting the power consumed<sup>141</sup> by a program is a non-trivial problem.

Programming languages are starting to support constructs that provide developers with options for dealing with power consumption issues.<sup>1198</sup>

?

### 13.4.2.1 Hardware variation

This section outlines the evidence for large variations in hardware performance. Much of the detailed data in the following analysis was obtained using components manufactured five to ten years before this book was published. The major computing hardware components include:

- CPU: power consumption and instruction counts,
- Main storage: hard disc performance and power consumption,
- Memory: performance and power consumption,
- Network performance,
- Aging of components.

The Intel Haswell processor is the latest iteration (at the time of writing) of the growth in processor variability.<sup>569</sup> This processor contains integrated power regulators per core, allowing each core to operate at a different voltage and frequency. The temperature of each core is monitored to ensure it keeps within thermal constraints, with voltage/frequency reductions when necessary. Use of the AVX instructions requires higher voltages and the processor automatically reduces clock frequency to keep within thermal limits; return to non-AVX operating mode takes 1 ms and I'm sure that in the coming years research will uncover some surprising consequences of this delay.

When computing devices obtain their electricity from the mains power supply, there is rarely a need to be interested in the supply characteristics. Batteries have characteristics that can

<sup>xii</sup> The following two sections separately discuss performance variation whose root cause in hardware or software; this is for simplicity of presentation.

affect the performance of devices connected to them, such as level of power delivery depending on current charge state and power draw frequency characteristics. In a mobile computing environment power consumption can be just as important, if not more so, than performance. Peltonen et al<sup>1073</sup> is a public dataset of power consumption on 149,788 mobile devices made up of 2535 different Android models; Jongerden<sup>708</sup> analyses various models of battery powered systems and Buchmann<sup>194</sup> covers rechargeable batteries in detail.

In mobile devices a large percentage of power is consumed by the display; optimization display intensity and choice of color,<sup>1297</sup> while an app is running, is not discussed here.

CMOS (complementary metal-oxide-semiconductor) is the dominant technology used to fabricate computing devices, and until a so-called beyond-CMOS device<sup>1011</sup> technology becomes commercially viable, only the characteristics of CMOS devices are considered.

**CPU** Developers have generally considered the processor executing their code to be interchangeable with any of the other mass-produced etched slices of silicon stamped with the same model number; while never exactly true, deviations from this interchangeability assumption were once small enough to only be of interest within specialised niches, such as hardware modders interested in running systems beyond rated limits e.g., higher clock rates.

The micro-architecture of modern processors has become so complicated that apparently minor changes to an instruction sequence can have a major impact on performance;<sup>652</sup> a trivial change to the source code or the use of a different compiler flag is enough.

Power consumption and clock frequency are intimately connected processor characteristics. Increasing clock frequency increases power consumption (a good approximation for processor power consumption is  $P = \alpha FV^2 + I_0V$ , where  $F$  is the clock frequency,  $V$  is voltage supplied to the cpu and  $I_0$  is leakage current). Processors clocked at the same frequency execute instructions at the same rate. However, variations in the number of atoms implementing internal circuitry produces variations in power consumption. Some processors reach maximum operating temperature more quickly than others; to prevent overheating destroying the device, power consumption is reduced by reducing the clock rate. Different processors have different sustained performance rates because of differences in their power consumption characteristics.

Vogeler<sup>325</sup> is a readable technical discussion of modeling low level temperature/power relationships for the kind of processors used to run applications.

[data-rbook/Exynos-7420/...](#)

?

?

A study by Wanner, Apte, Balani, Gupta and Srivastava<sup>1426</sup> measured the power consumed by 10 separate Amtel SAM3U microcontrollers (derived from an ARM's Cortex M3 processor core) at various ambient temperatures. Figure 13.18 shows a 5-to-1 difference, between supposedly identical processors, in power consumption when in sleep-mode and around 5% difference when operating at 4MHz.

Another example of power variations, involving the Intel Core processor, is discussed in Section 11.6 on building mixed-effects models.

Any power related benchmark made using a single instance of a processor is a sample drawn from a population that could vary by 10% or more when executing code and several hundred percent when idling. The extent to which results based on this minimum sample is of practical use will depend on the questions being asked. If the power consumption characteristics of the population of a particular CPU is required, then it is necessary to benchmark a sample containing an appropriate number of *identical* processors.

Power consumption measurements are often implemented by periodically sampling the voltage across a known resistance. A study by Saborido, Arnaoudova, Beltrame, Khomh and Antoniol<sup>1189</sup> investigated the measurement error introduced by different sampling rates, on mobile devices. Figure 13.19 shows the power spectrum of the Botanica App, executing on a BeagleBone Black running Android 4.2.2, sampled at 500K per second. This very high sampling rate makes it possible to see the noticeable peak in power consumed by very short-lived events, something that low frequency sampling would not detect (the paper lists error estimates for lower sampling rates).

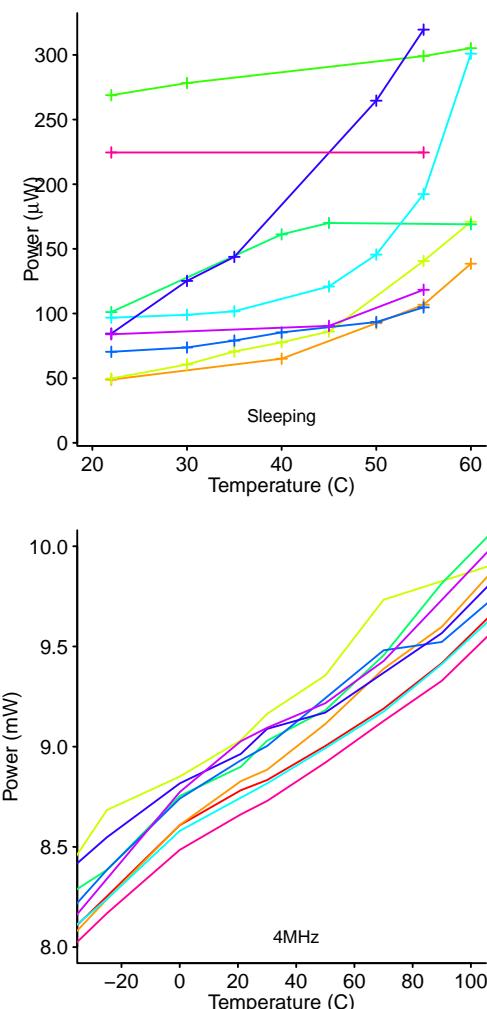


Figure 13.18: Power consumed by 10 Amtel SAM3U microcontrollers at various temperatures when sleeping or running. Data from Wanner et al.<sup>1426</sup> [code](#)

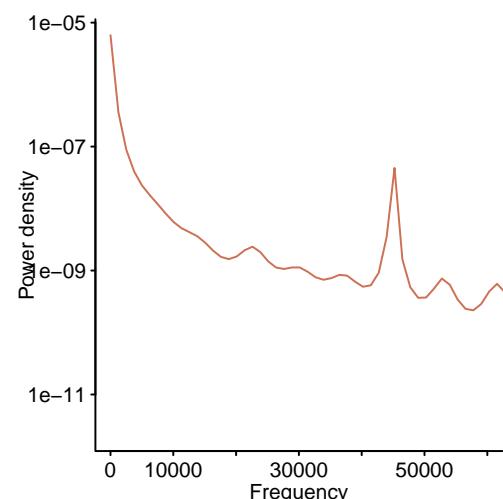


Figure 13.19: Power spectrum of electrical power consumed by the Botanica App executing on a BeagleBone Black running Android 4.2.2. Data from Saborido et al.<sup>1189</sup> [code](#)

In theory counting the instruction executed by a program is a means of obtaining, power independent, answers to questions about comparative program performance. Many processors include hardware containing counts of operations performed, e.g., instruction opcodes executed and cache misses; however, the purpose of these counters is to help manufacturers debug their processors, not provide end-user functionality. Consequently, counter values are not guaranteed to be consistent across variants of processors within the same family<sup>1433</sup> and fixing faults in the counting hardware does not have a high priority (e.g., counting some instructions twice or not at all; Weaver and Dongarra<sup>1433</sup> found that in most cases the differences were a fraction of a percent of the total count, but for some kinds of instructions, such as floating-point, the counts were substantially different).

A study by Melhus and Jensen<sup>932</sup> showed that address aliasing, of objects in memory, could have a huge impact on the relative values of some hardware performance counters.

Counting the number of instructions executed by a program begs the question of how calls into operating system routines, which may execute at higher privilege levels, are counted. Also, the execution time of some instructions will depend on other instructions executed at around the same time (modern processors have multiple functional units and allocate resources based on the instructions currently in the pipeline), the time taken to execute other instructions can be very unpredictable (e.g., time taken to load a value from memory depends on the current contents of the cache and other outstanding load requests). A study by Weaver and McKee<sup>1434</sup> found that it was possible to adjust for the known faults in the hardware counters (see [benchmark/iiswc2008-i686.R](#)).

Hardware counters need not be immune to *observer effects*; in particular, the values returned can depend on the number of different hardware counters being collected.<sup>985</sup>

**Main storage** Traditionally main storage has meant hard disks (sometimes backed-up with tape), but solid state devices (SSD) are rapidly growing in capacity<sup>1328</sup> and importance; for extremely large capacity magnetic tape is used: this niche use is not discussed here.

Data is read/written to a hard disk by moving a magnetic sensor across a rotating surface. These spatial movements create a correlation between successive operations, e.g., the time taken to perform the second read will depend on its location on the disk relative to the first. Disk access issues have been studied for around 50 years and techniques such as data buffering, by the operating system, and reordering requests to optimise overall throughput, by the device driver or firmware, are well-established.<sup>?</sup>

Figure 13.20 shows that data near the outside of one (not unusual) disk family is read approximately twice as fast as data located near the center. This offset dependent performance has always existed in some form and its impact on benchmark performance is very difficult to estimate, depending on the history of the data that is added and deleted.

Storage farms organise files so that those most likely to be accessed are stored on the outer tracks, while files less likely to be accessed are stored on the inner tracks.

The continuing increase in the number of bits that can be stored within the same area of rotating rust has been achieved by reducing the size of the magnetic domain used to store a bit. Like silicon wafer production, variations in the fabrication process of disc platters can now result in large differences in the performance of supposedly identical drives. A growing percentage of disks are used in data centers and at some point manufacturers may decide to concentrate on designing drives for this market.<sup>180</sup>

A study by Krevat, Tucek and Ganger<sup>786</sup> measured the performance of disk drives originally sold in 2002, 2006, 2008 and 2009. In Figure 13.20 the staircase effect is a result of zoning? (disks spin at a constant rate and in the same time interval more data can be read from the area swept out near the outer edge of a platter than from one near the inner edge).

The upper plot shows the performance of nine disks from 2002, each displayed using a different color and there is little variation between different disks (fitting a regression models shows that disk identity is not a significant, p-values around 0.2, predictor of performance).

The lower plot is nine disks from 2006, each displayed using a different color and the visibility of different colors shows there is a noticeable variation between different disks (fitting a regression models shows that disk identity is a significant, p-values around  $10^{-16}$ , component of performance prediction).

To increase recording density, drive manufacturers are now using Shingled Magnetic Recording (SMR), where tracks overlap like rows of shingles on a roof. Singled discs have very different performance characteristics,<sup>11</sup> but little data is publicly available at the time of writing.

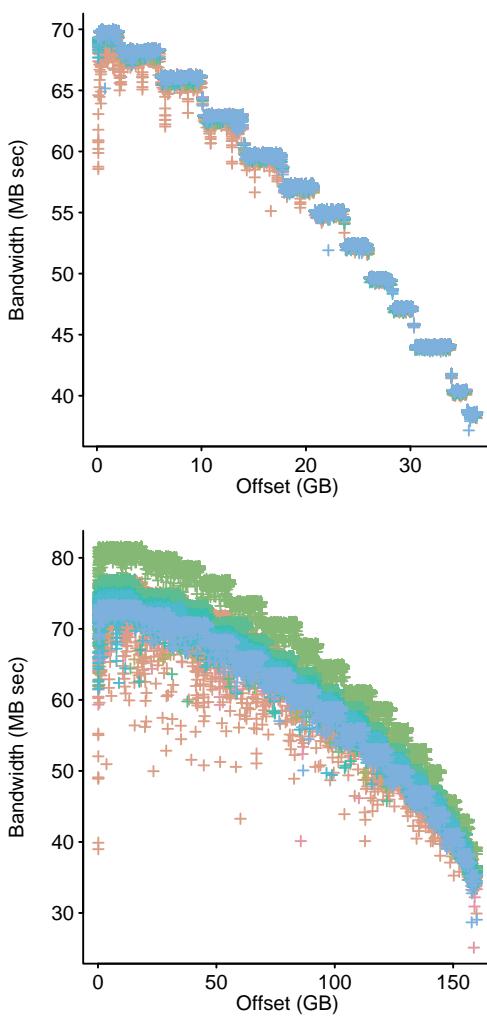


Figure 13.20: Read bandwidth at various offsets for new disks sold in 2002 (upper) and 2006 (lower). Data kindly provided by Krevat.<sup>786</sup> [code](#)

SSDs are sufficiently new that little performance data is publicly available at the time of writing.

A study by Kim<sup>752</sup> ran eight different benchmarks on SSD cards from nine different vendors. The range of performance values was different for both vendors and benchmark, meaning that in their original form neither of these variables are of any use in modeling performance. However, if the performance on each benchmark is normalised across all vendors (e.g., by transforming into the range zero to one) it might be possible to fit a model using vendor as the single explanatory variable (see [benchmark/hyojun/hyojun.R](#)). The results show that only one vendor has a sufficiently consistent performance (that appears to be significantly better than the other vendors) to be included in a model, with all other vendors appearing to have the same performance.

**Memory** Memory chips tend to be thought about in terms of their capacity and not their performance (e.g., read/write delays or power consumption). Performance is governed by access rate and by the number of bytes transferred per access, with accesses usually made via some form of memory control chip (the capabilities of this controller have a significant impact on performance). Many motherboards provide options to select memory chip timing characteristics.

The upper plot in Figure 13.21 breaks down average power by CPU and memory when running the SPEC CPU2006 benchmark, while the lower plot breaks the power down by the major subcomponents of a server running various programs.

There is a storage performance hierarchy, with on (cpu) chip cache providing faster access to frequently used data. The interaction between the size of the various storage caches and an algorithm's memory can result in performance characteristics that change as the size of objects processed grows.

A study by Khuong and Morin<sup>749</sup> measured the performance of several search algorithms when operating on arrays of various sizes (items were stored appropriately in the array, for the algorithm used). Figure 13.22 shows the time taken by different search algorithms to find an item in an array, for arrays of various sizes; the grey lines show the total size of the processor L1, L2 and L3 caches.

The following are examples of memory chip characteristics that have been found to noticeably variation:

- a study by Gottscho, Kagalwalla and Gupta<sup>527</sup> measured power consumption variability of 13 DIMMs of the same model of 1G DRAM from four vendors. The variation about the mean, at one standard deviation, was 5% for read operations, 9% for write and 7% for idling (see [benchmark/J20\\_paper.R](#)),
- a study by Schöne, Hackenberg and Molka<sup>1210</sup> found that memory bandwidth was reduced by up to 60% as the frequency of the cpu was reduced and that memory performance characteristics varied between consecutive generations of Intel processors and even between server and desktop parts,
- a study by Gottscho<sup>526</sup> measured the power consumption of 22 DDR3 DRAMs, manufactured in 2010 and 2011, from four vendors. Read operations consumed around 60% of the power needed for write operations, with idle consuming around 40%; the standard deviation varied from 10% to 20%. The power consumed also varied with value being read/written, e.g., writing 1 to storage containing a 0 required 25% more power than writing a 0 over a 1 (see [benchmark/MSTR10-DIMM.R](#) for data).

The variability of memory chip performance is likely to increase as vendors start to reduce power consumption and improve performance by lengthening DRAM refresh times, optimising each computer by tuning it to the unique characteristics of the particular chips present in each system.<sup>827</sup>

Chandrasekar<sup>231</sup> provides a detailed discussion of DRAM power issues and includes code for a tool to obtain detailed information about the memory chips installed on a system.

**Network performance** A study by Schad, Dittrich and Quiané-Ruiz<sup>1204</sup> submitted various benchmarks as jobs to Amazon's Elastic Computing...

**Aging** emailed for data...

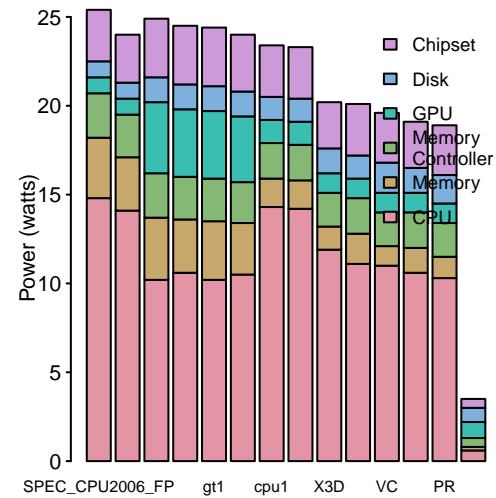
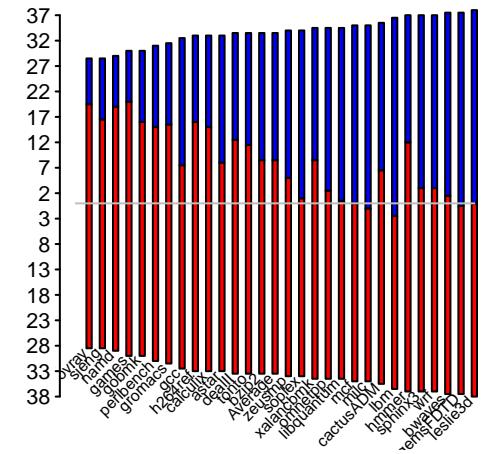


Figure 13.21: Average power consumed by one server's CPU (four Pentium 4 Xeons; red) and memory (8 GB PC133 DIMMs; blue) running the SPEC CPU2006 benchmark (upper) and breakdown by system component when executing various programs. Data from Bircher.<sup>141</sup> [code](#)

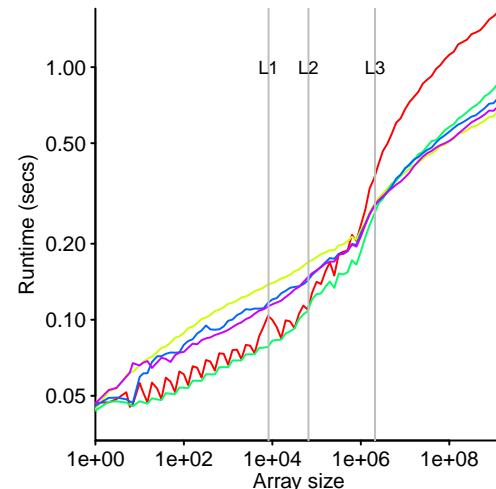


Figure 13.22: Time taken to find a unique item in arrays of various size, containing distinct items, using various search algorithms; grey lines are L1, L2 and L3 processor cache sizes. Data from Khuong et al.<sup>749</sup> [code](#)

### 13.4.2.2 Software variation

This section outlines the evidence for large variations in software performance. The following software characteristics are briefly covered:

- The environment: interaction with the environment, file system, support libraries and aging,
- Configurations,
- Creating an executable: compiler optimization and link order,
- Tools.

**The environment** The environment within which programs execute often contains a complicated ensemble of interconnecting processes and services that cannot be treated as independent standalone components. One consequence of this complexity<sup>696</sup> and interconnectedness is that the order in which processes are initiated during system startup can have a noticeable impact on system performance.

The impact of prior history on program performance is seen in a study by Kalibera, Bulej and Tůma<sup>725</sup> who measured the execution time of various programs. Figure 13.23 shows 10 iterations of the procedure: reboot computer and make 2,048 performance measurements. The results show performance variation after each reboot is around 0.1%, but rebooting can cause a shift of 3% in the average performance (the ordering of processes executed during system startup varies across reboots, due to small changes in the time taken to execute the many small scripts that are invoked during startup).

A later study<sup>624</sup> found that the non-determinism of initial program execution, in this case, could be reduced by having the operating system use cache-aware page allocation.

Environmental interactions are not always obvious. A study by Mytkowicz, Diwan, Hauswirth and Sweeney<sup>985</sup> increased the number of bytes occupied by a Linux environment variable between runs of the perlbench program; the results from each of 15 executions were recorded, an environment variable increased in size by one character and this process repeated 100 times.

Figure 13.24 shows the percentage change in performance, relative to no environment variables, in perlbench performance as characters are added to the environment. Data extracted from Mytkowicz et al.<sup>985</sup> code

Incremental operating system updates can change program performance. A study by Flater<sup>444</sup> compared the performance of cpu intensive and I/O bound programs on two different versions of Slackware running on the same hardware (versions 14.0 and 14.1, using Linux kernels 3.12.6 and 3.14.3 respectively). The results show consistent differences in performances of up to 1.5% (rebooting did not have any significant impact on performance).

Many systems allow multiple programs to be executing at the same time, sharing system resources. Sharing becomes a performance bottleneck when one program cannot immediately access resources when it requests them; access to memory is a common resource contention issue on multi-processing systems.

Figure 13.25 shows changes in SPEC CPU2006 benchmark performance caused by cache and memory bus contention on a dual processor Intel Xeon E5345 system (from a study by Babka<sup>70</sup>).

A study by Mazouz<sup>907</sup> investigated the performance of the SPEC OpenMP 2001 benchmark programs, compiled using gcc 4.3.2 and icc 11.0, running on multicore devices. It is possible for a program's code to execute on a different core after every context switch. Allowing the operating system to select the core to continue program execution is good for system level load balancing, but can reduce the performance of individual programs because recently accessed data is less likely to be present in the cache of the newly selected core. *Thread affinity* is the process of assigning each thread to a subset of cores, with the intent of improving data locality i.e., recently accessed data is more likely to be available in accessible caches.

Figure 13.26 shows the time taken to execute one program in 2, 4, and 6 threads, with thread affinity set to scatter (distribute the threads evenly over all cores), no affinity (allow the OS to assign threads to cores) and compact (threads share an L2 cache), each repeated 35 times.

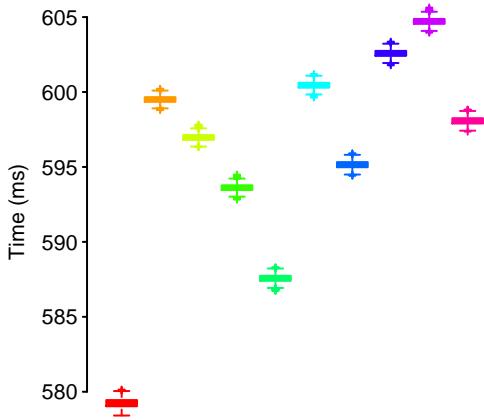


Figure 13.23: FFT benchmark executed 2,048 times followed by system reboot, repeated 10 times. Data kindly provided by from.<sup>725</sup> code

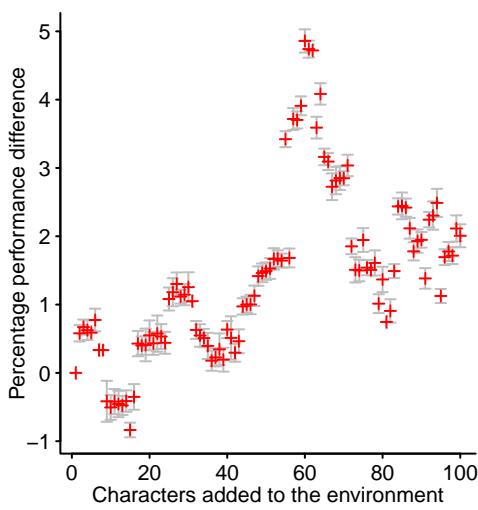


Figure 13.24: Percentage change, relative to no environment variables, in perlbench performance as characters are added to the environment. Data extracted from Mytkowicz et al.<sup>985</sup> code

- shared cache: Hits
- shared cache: Conflict misses
- shared cache: Capacity misses
- shared memory bus: Accesses

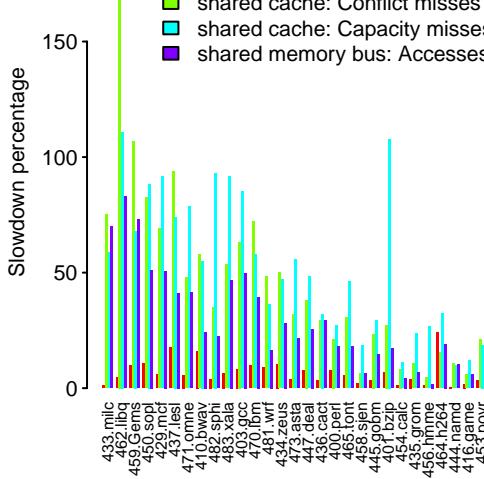
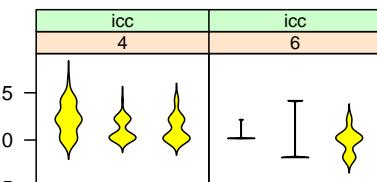


Figure 13.25: Changes in SPEC CPU2006 benchmark performance caused by cache and memory bus contention for one dual processor Intel Xeon E5345 system. Data kindly provided by Babka.<sup>70</sup> code



Configuring the system being benchmarked to only run one program at a time solves some, but not all, cache contention issues. Walking through memory, in a loop, may result in a small subset of the available cache storage being used (main memory is mapped to a much smaller cache memory, which means that many main memory addresses are mapped to the same cache address). Figure 13.27, from a study by Babka and Tåma,<sup>71</sup> shows the effect of walking through memory using three different fixed width strides; for 32 and 64 byte strides accesses to even cache lines is faster than odd lines, with the pattern reversed for a 128 byte stride.

Operating systems generally have background processes that spend most of the time idling, but wake up every now and again. A background processes that wakes up will consume system resources and can have an impact on the performance reported by a benchmark, they are a source of variation.

A study by Larres<sup>818</sup> investigated how the performance of one version of Firefox changed as various operating system features were disabled (the intent being to reduce the likelihood that external factors added noise to the result). The operating system features modified were: 1) every process that was not necessary was terminated, 2) address-space randomization was disabled, 3) the Firefox process was bound exclusively to one cpu, and 4) the Firefox binary was copied to and executed from a RAMDISK.

The Talos benchmark was used (the performance testing framework used by Mozilla) and every program was run 30 times. Figure 13.28 shows the performance of various programs running in original and stabilised (i.e., low-noise) configurations.

Those users interested in consistent performance will want to minimise the variation in benchmark results (which did occur for some programs), while users interested in actual benchmark performance will be interested that significant changes in the mean occurred for some programs.

**File systems** File systems are a way of organising information on a storage device. The traditional view of a file being just a leaf in a directory tree has become blurred, with many file system managers now treating compressed archived files (e.g., zip files) as-if they had a directory structure that can be traversed and Microsoft's .doc format containing a FAT (File Allocation Table, just like a mounted Windows file system) that can refer to contents that may be distributed outside of the file.

A study by Zhou, Huang, Li and Wang<sup>1496</sup> looked at the performance interplay between file systems and Solid State Disks (SSD) by running a file-server benchmark on a Kingston MLC 60GB SSD. Four commonly used Linux filesystems (ext2, ext3, reiserfs and xfs) were mounted in turn using various options, e.g., various block sizes, noatime, etc.

Figure 13.29 shows the number of operations per second for a file-server benchmark (see paper and data for other benchmarks). The data is poorly fitted by a linear regression model (i.e., not significant on the filesystem or mount options; see [benchmark/filesystem-SSD.R](#)).

A study by Sehgal, Tarasov and Zadok<sup>1223</sup> compared the power used when four commonly used filesystems were mounted in various ways, e.g., fixed vs. variable sector size, different journal modes, etc. Various server workloads running on Linux were measured; web server power consumption varied by a factor of eight, mail server by a factor of six and file and database by a factor of two.

**Creating an executable** Many applications are built by translating source code to an executable binary, with the translation tools often supporting many options, e.g., gcc supports over 160 different options for controlling machine independent optimization behavior. Compiler writers strive to improve the quality of generated code and it is to be expected that the performance of each release of a compiler will be different from the previous one; there have been around 150 released versions of gcc in its 30 year history.

A study by Makarov<sup>889</sup> measured the performance of nine releases of gcc made between 2003 and 2010, on the same computer using the same benchmark suite (SPEC2000), at optimization levels O2 and O3.

Figure 13.30 shows the percentage change in SPEC number, relative to version 4.0.4, for the 12 integer benchmark programs compiled using six different versions of gcc. SPEC has a long history if being used for compiler benchmarking and it is possible that all the versions of gcc used for this comparison have already been tuned to do well on this benchmark, meaning there is little, benchmark specific, improvement to be had in the successive versions used in this study.

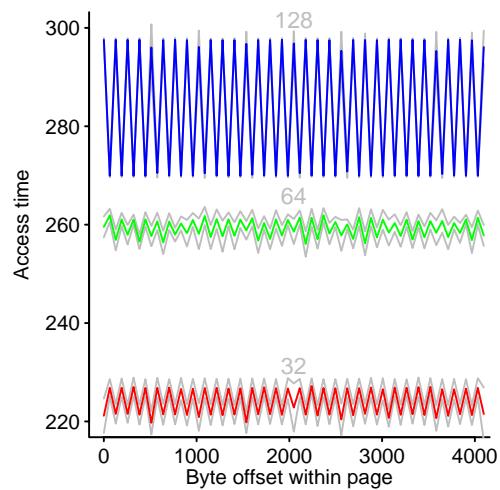


Figure 13.27: Access times when walking through memory using three fixed stride patterns (i.e., 32, 64 and 128 bytes) on a quad-core Intel Xeon E5345; grey lines at one standard deviation. Data kindly provided by Babka.<sup>71</sup> [code](#)

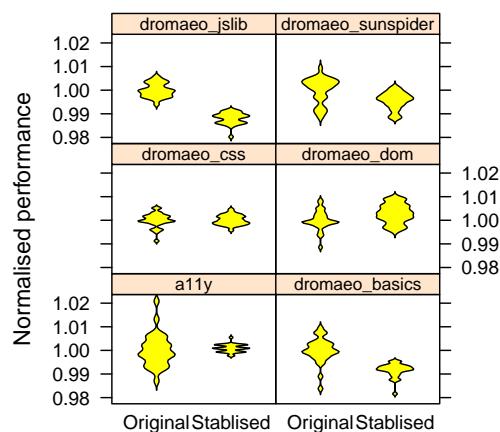


Figure 13.28: Performance variation of programs from the Talos benchmark run on original OS and a stabilised OS. Data from Larres.<sup>818</sup> [code](#)

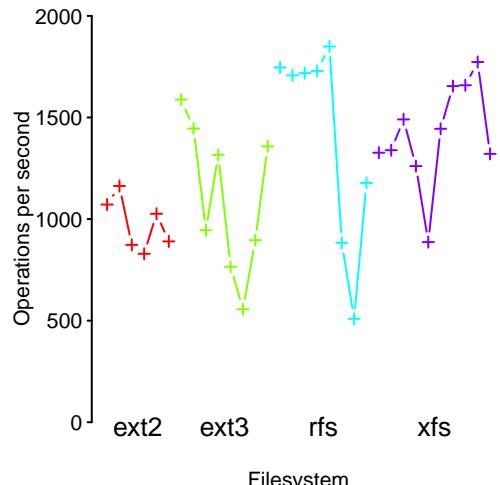


Figure 13.29: Operations per second of a file-server mounted on one of ext2, ext3, rfs and xfs filesystems (same color for each filesystem) using various options. Data kindly supplied by Huang.<sup>1496</sup> [code](#)

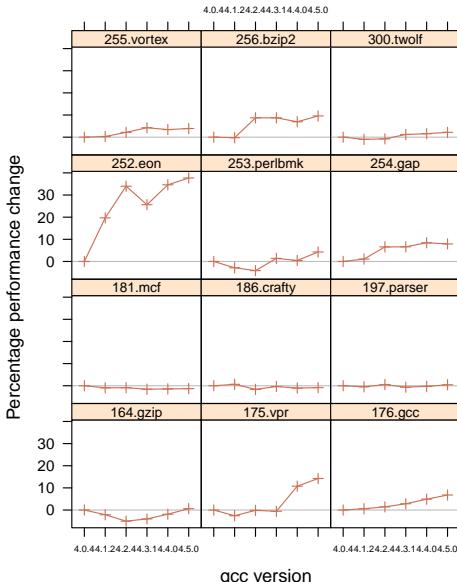


Figure 13.30: Percentage change in SPEC number, relative to version 4.0.4, for 12 programs compiled using six different versions of gcc (compiling to 64-bits with the -O3 option). Data from Makarov.<sup>889</sup> code

The following summary output is from a mixed-effect model with the random effect on the intercept and slope: `code`

```
Linear mixed model fit by REML [‘lmerMod’]
Formula: value ~ gcc_version + (gcc_version | Name)
Data: lme_02
```

REML criterion at convergence: 400.6

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.7256	-0.2748	-0.0683	0.3039	4.3372

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
Name	(Intercept)	1192.795	34.537	
	gcc_version	3.155	1.776	-1.00
Residual		8.632	2.938	

Number of obs: 72, groups: Name, 12

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	-29.7469	11.0553	-2.691
gcc_version	1.4126	0.5513	2.562

Correlation of Fixed Effects:

(Intr)
gcc_version -0.997

The general picture paint by the model results is a small improvement with each gcc release is swamped by the size of the random effects, while the picture pained by Figure 13.30 is of some releases having a large impact on some programs.

A study by de Oliveira, Petkovich, Reidemeister and Fischmeister<sup>323</sup> investigated the impact of compiler optimization and object module link order on program performance. Figure 13.31 shows the time taken by the xy file compression program, compiled by gcc using various optimization options, to process the Maximum Compression test set on various systems. The results show that different optimization levels have a different performance impact on different systems (the lines would be parallel if optimization level had the same impact for each system).

Compiling is the first step in the chain of introducing variability into program performance, the next step is linking. Figure 13.32 shows execution time of Perlbench (one of the SPEC benchmark programs), on six systems, when the object files used to build the executable are linked in three different orders and with address randomization on/off.

Some systems share a consistent performance pattern across link orderings and some systems are not affected by address randomization. But there is plenty of variation across all the variables measured.

**Tools** Dynamic profiling tools such a grprof work by interrupting a program at regular intervals during execution (e.g., once every 0.01 seconds) and recording the current code location (often at the granularity of a complete function). The results obtained can depend on interrupt frequency and the likelihood of being in the process of calling/returning from the profiled function.<sup>443</sup>

### 13.4.2.3 End user systems

Benchmark data derived from end user systems is likely to be subject to numerous known and unknown unknowns.

PassMark Software specializes in benchmark solutions<sup>1465</sup> and over the years has collected 800,000+ benchmark results from Microsoft Windows based computers; David Wren kindly supplied the 10,000 memory benchmark results used in the following analysis, as well as insights into possible reasons for the performance characteristics seen.

The results obtained from end users running a benchmark on one or more of their computers will depend on the characteristics of the hardware and the software executing at the time the

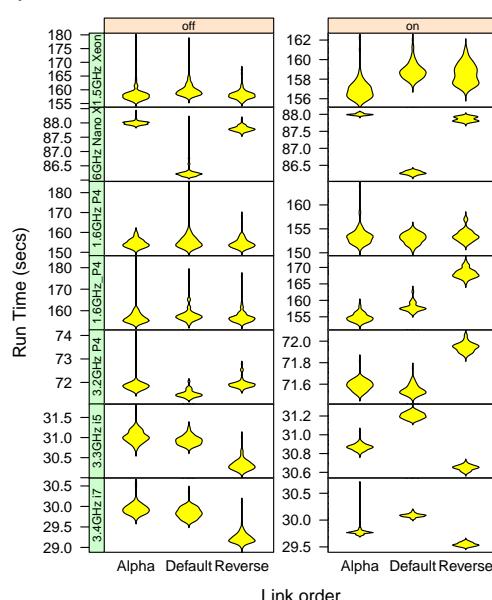


Figure 13.31: Execution time of xy file compressor, compiled using gcc using various optimization options, running on various systems (lines are mean execution time when compiled using each option). Data kindly supplied by Petkovich.<sup>323</sup> code

v 0.10.0

Figure 13.32: Execution time of Perlbench, from SPEC benchmark, on six systems, when linked in three differ-

benchmark is run. Background processes that may be running on a Windows machine include Internet based toolbars, anti-virus systems and general OS housekeeping processes.

The upper plot in Figure 13.33 shows the results (in ascending order) for 783 systems containing an Intel Core i7-3770K processor (whose official clock speed is 3.5GHz, some users may be overclocking); the lower plot has had the values at each end trimmed by around 10% and the red dots are predictions from a regression model built using information on the characteristics of the memory chips in each computer as explanatory variables.

The scattering of red dots around the regression line illustrates how poor the predictions can be from a regression model that explains about 30% of the variance in the data. This data is another example (see Figure 13.9) of the wide range of performance reported for apparently very similar end user systems.

#### 13.4.2.4 The cloud

Cloud computing is becoming a popular platform for running applications that require non-trivial compute resources. The service level agreements offered by cloud providers specify minimum levels of service, e.g., Amazon's June 2013 EC2 terms specify 99.95% monthly uptime.<sup>31</sup> Cloud services generally run virtualized instances, which means access to the real hardware may sometimes be shared. Shared hardware access results in user visible performance varying from one run to the next; what are the characteristics of this variation?

A study by Schad, Dittrich and Quiané-Ruiz<sup>1204</sup> submitted various benchmarks as jobs to Amazon's Elastic Computing Cloud (EC2) twice an hour over a 31-day period; a variety of resource usage measurements were recorded. Figure 13.34 shows one set of resource usage measurements from the study, the Unix benchmark utility (Ubench; a cpu benchmark) running on small (upper) and large (lower) EC2 instances located both in Europe (red) and the US (green).

Both plots show more than one distinct ranges of performance. This data is an example of the variation experienced in Amazon's EC2 performance over one particular time period and there is no reason to believe that any subsequent benchmarking will exhibit one, two, three or more distinct performance ranges.

## 13.5 User interface testing

When working on systems being built by a small group of people, software developers are called on to solve a wide range of computer related issues, including the testing and tuning of the user interface.

The System Usability Scale (SUS)<sup>187,188</sup> is a straight-forward and widely used usability questionnaire that produces a single number for usability. One study<sup>1368</sup> comparing five methods of evaluating website usability found that SUS produced the most consistent results for smaller sample sizes.

User interface design...?

## 13.6 Surveys

A lot of software engineering information only exists in the heads' of the people who build software systems. Obtaining this information requires asking these people questions and analysing their answers. This is the subject of survey analysis.

A survey is essentially an experiment where the questions are used to control the explanatory variables.

The choice of the population to sample, for a survey, should be driven by where accurate answers to the questions are most likely to be obtained; practical considerations may dictate the use of alternative populations.

The survey package ...

The techniques used to analyse survey results are the same as those used to analyse other kinds of data. However, there are some characteristics that are often encountered in survey data, including the following:

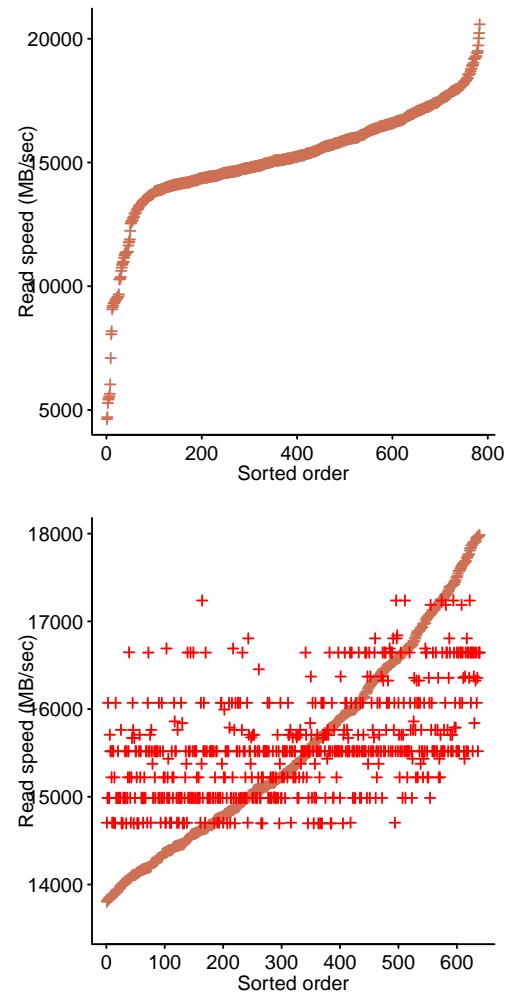
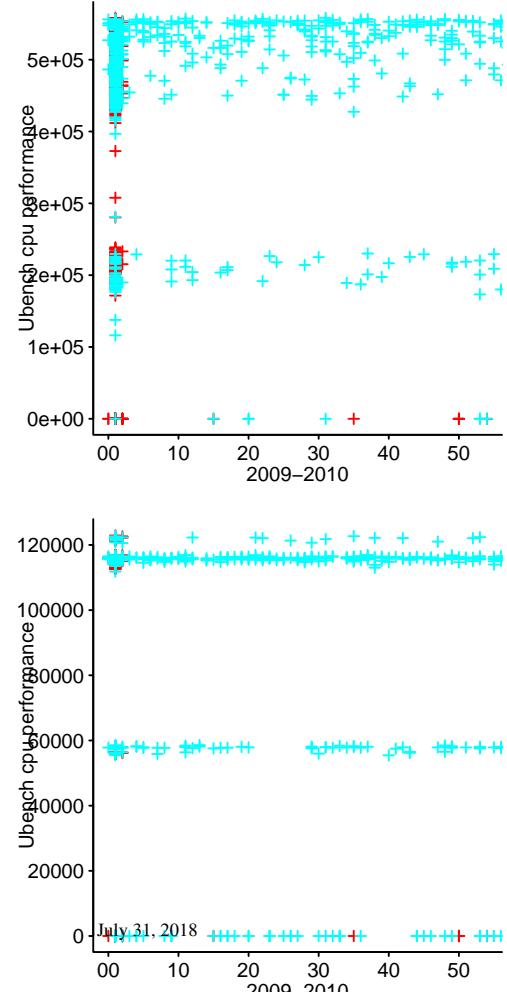


Figure 13.33: Performance of PassMark memory benchmark on 783 Intel Core i7-3770K systems; lower plot created by trimming 10% of values from the ends of what appears in the upper plot. Data kindly supplied by David Wren.<sup>1465</sup> code



- missing data: people don't answer all the questions or stop answering after some point,
- misleading answers: giving answers that show those involved in a better light, such as job adverts listing trendy topics and languages to attract more applicants,
- spatial information: how subjects are distributed geographically,
- stratification: to increase the number of responses the survey is sent to people with a wider range of characteristics than is strictly applicable, potentially resulting in samples containing clusters having their own distinct characteristics,

Like most experiments, finding the appropriate questions to ask is an iterative process. It is worthwhile running test surveys, to obtain feedback on whether the questions produce the desired kind of response, updating the questions based on this feedback and repeating as many often as time and resources allow.

Studies have found<sup>368</sup> that self-assessment of skills and character have a tenuous to modest relationship with actual performance and behavior. The correlation between self-ratings of skill and actual performance in many domains is moderate to meager....

Several studies by your author<sup>700, 701, 703</sup> included a component that asked developers about how many lines they had read and written during their professional career.

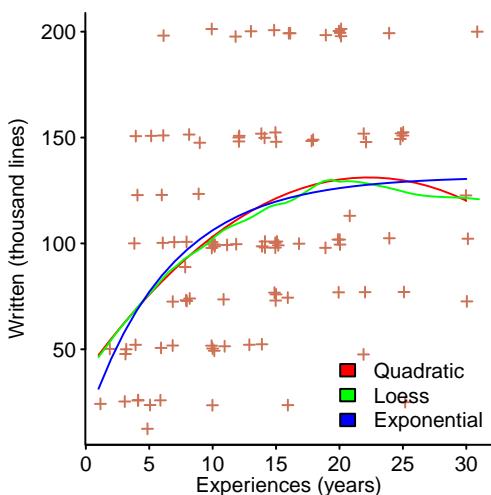


Figure 13.35: Number of lines of code that 101 professional developers, with a given number of years experience, estimate they have written, lines are various regression fits. Data from Jones.<sup>700,701,703</sup> code

This question requires a lot of thought to answer and there are many ways of adding up the numbers. Does reading the same line twice count as two lines, or one line unless the developer involved had forgotten reading it? How much does visually searching a screen of code for a particular identifier count towards lines read? Counting the number of lines in the programs written by a developer is likely to underestimate the number of lines they have written; a line of code may be written and then deleted, an existing line may be modified slightly.

Figure 13.35 shows the number of lines of code that 101 professional developers estimate they have written. While an exponential model may fit the data, the variance explained is small...

How well do developers' know a particular language? Answering this question first requires a definition of what it means to know a language. A study by Dietrich, Jezek and Brada<sup>353</sup> investigated one aspect of knowing a language: knowledge of an important component of the language semantics, in this case Java type compatibility.

They started with a thorough set of questions, but quickly found that subjects were not answering many questions; they switched to a shorter list of questions.

As a self administered survey, they had no control over how many questions were answered...

The questions required detailed technical knowledge... breaking down what was asked into subcomponents might have been better than an all or nothing approach to the correct answer...

Allowing subjects to provide more specific values for experience...

#### developers/java-type-quest.R

Some surveys are made to gain general information about the characteristics of a particular population of interest. The questions are designed to learn about the characteristics of this population, e.g., characteristics of open source developers,<sup>1169</sup> such as the age when a person started contributing to open source projects, see Figure 8.12.

The results of many studies<sup>44</sup> have found that most subject ratings are based on an ordinal scale (i.e., there is no fixed relationship between the difference between a rating of 2 and 3 and a rating of 3 and 4), that some subjects will be overly generous or miserly in their rating and that without strict rating guidelines different subjects apply different criteria when making their judgements (which can result a subject providing a list of ratings that is inconsistent with every other subject).

the TopKLists package implements various algorithms... For this analysis a simple aggregate ranking algorithm using Borda's method<sup>852</sup> was implemented.

145 Questions for data scientists in software engineering<sup>117</sup>

How often hardware and software is updated, a basic survey<sup>1350</sup>...

What questions to ask?... ?, ?, ?

Questionnaire data on training and related issues... Difference in survey answers between years...

2008 answers:<sup>?</sup>

2009 answers:<sup>?</sup>

[embedded\\_survey/](#)

### 13.6.1 Checking survey reports

Does mean age make sense given the sample size..

Solving an underspecified set of equations... Survey with only totals given [Success2007.pdf](#)...



# Chapter 14

## Overview of R

This chapter gives a brief overview of R for developers who are fluent in at least one other computer language. The discussion pays attention to language features are very different from languages the reader is likely to be familiar with and concentrate on a few language features that can be used to solve most problems.

The R language is defined by its one implementation; available from the R core team.<sup>1135</sup> A language definition<sup>1136</sup> is gradually being written.

R programs tend to be very short, compared to programs in languages such as C++ and Java; 100 lines is a long R program. It is assumed that most readers will be casual users of R whose programs generally follow the pattern:

```
d=read_data()  
clean_d=clean_and_format(d)  
d_result=applicable_statistical_routine(clean_d)  
display_results(d_result)
```

If your problem cannot be solved using this algorithm, then perhaps the most efficient solution may be for you, dear reader, to use the languages and tools you are already familiar with to preprocess the data so that it can be analysed and processed using R.

R is a domain specific language whose designers have done an excellent job of creating a tool suited to the tasks frequently performed when analysing the kinds of datasets encountered in statistical analysis. Yes, R is Turing complete, so any algorithm that can be implemented in other programming languages can be implemented in R, but it is designed to do certain things very well with no regard to making it suitable for general programming tasks.

As a language the syntax and semantics of R is a lot smaller than many other languages. However, it has a very large base library, containing over 1,000 functions. Most of the investment needed to become a proficient R user needs to go into learning to use and apply these functions. There are over 6,000 add-on packages available from the CRAN (Comprehensive R Archive Network).

Help on a specific identifier, if any is available, can be obtained using the ? (question mark) unary operator followed by an identifier. The ?? unary operator, followed by an identifier, returns a list of names associated with that identifier for which a help page is available.

The call `library(help=circular)` lists the functions and objects provided by the package named in the argument.

### 14.1 Your first R program

Much like in Python, Perl and many other interpreted implementations, R can be run in an interactive mode, where code can be typed and immediately executed (with "Hello world" producing the obvious output).

Your first R program has to read some data and plot it, not just print "Hello World". The following program reads a file containing a single column of values and plots them:

```
the_data=read.csv("hello_world.csv")
plot(the_data)
```

to produce Figure 14.1.

The `read.csv` function is included in the library that comes bundled with the base system (functions not included in this library have to be loaded using the `library` function before they can be referenced; they may also need to be installed via the `install.packages` function) and has a variety of optional arguments (arguments can be omitted if the function definition includes default values for the corresponding parameter). Perhaps the most commonly used optional arguments are `sep` (the character used to separate values on a line, defaults to comma) and `header` (whether the contents of the first line should be treated as column names, default TRUE).

The value returned by `read.csv` has class `data.frame`, which might be thought of as a C struct type (it contains data only, there are no member functions as such).

The `plot` function attempts to produce a reasonable looking graphic of whatever data is passed, which for character data is a histogram of the number of occurrences. R is not intended for manipulating low level details and some work is needed to get at the numeric values of the characters which appear in the second plot.

There are a wide variety of options to change the appearance of plot output; these can be applied on each call to `plot` or globally for every call (using the `par` function).

All objects in the current environment can be saved to a file using the `save` function and a previously saved environment can be restored using the `load` function. When quitting an R session the user is given the option of saving the current environment to a file named `.RData`; if a file of this name exists when R is started, its contents are automatically loaded.

## 14.2 Language overview

R is a language and an environment. Like Perl, it is defined by how its dominant implementation behaves (i.e., the software maintained by the R project<sup>[1135](#)</sup>).

R was designed, in the mid-1990s, to be largely compatible with S (a language, which like C, started life in the mid-1970s at Bell Labs). When S was created, Fortran was the dominant engineering language and the Fortran way of doing things had a strong impact on early design decisions, compared to the C way.

The designers of R have called it a functional language and it does support a way of doing things that is most strongly associated with functional program languages (including making life cumbersome for developers wanting to assign to global variables).

The language also contains constructs that are said to make it an object oriented language, and it certainly contains some features found in object oriented languages. OO constructs were first added in the third iteration of the S language and are more of an addition to the functional flavor of the language than a complete make-over. The primary OO feature usage is function overloading when accessing functions from library packages.

Lateral thinking is often required to code a problem in R, using knowledge of functions contained in the base system, e.g., calling `order` to map a vector of strings to a unique vector of numbers.

### 14.2.1 Differences between R and widely used languages

The following list describes language behaviors that are different from that encountered in other commonly used languages:

- There are no scalars, e.g., 2 is a vector containing one element and is equivalent to writing `c(2)`. The unary and binary operators operate on complete vectors (among other things). Many operations that involve iterating over scalar values in other languages, e.g., adding two arrays, can be performed without explicit iteration in R, e.g., `c(1, 2) + c(3, 4)` has the value `c(4, 6)`,

- arrays start at one, not zero,
- matrices and data frames are indexed in row-column order (C-like languages use column-row order),
- case is significant in identifiers, e.g., `some_data` and `Some_data` are considered to refer to different objects,
- some language constructs implemented via specific language syntax in other languages are implemented as function calls in R, e.g., the functionality of `return` and `switch` is provided by function calls,
- there is a special operator, `<-` to assign to a variable in an outer scope from within the current function,
- vectors/arrays/data.frames can be sliced to return a subset of the original,
- explicit support for NA (Not Available). This value denotes a number that exists, but whose value is unknown. Operations involving NA return NA when the result value is not known because the value of NA is unknown, but will return a value when the result is independent of the value of NA, e.g., `NA || TRUE`,
- type conversion behavior may be driven by semantics rather than the underlying representation, e.g., `as.numeric("1") == 1` and `as.numeric("a")` returns NA.

The following R language features are found in commonly used languages:

- objects and functions come into existence during program execution when they appear on the left-hand-side of an assignment, function parameter, or in more obscure ways, and a value is assigned (there is no mechanism for declaring any kind of identifier),
- the type of an object is the type of the value last assigned to it,
- decimal and hexadecimal literals have type numeric (literals starting with zero are not treated as octal literals; the zero is ignored) even if they look like integers, because they do not contain a decimal point). Some input functions, e.g., `read.csv`, will consider a column to have integer type if all its values can be represented as an integer,
- what most other languages considered to be a statement (i.e., something that does not return a value) R treats as an expression (e.g., `if/for` statements return a value).

## 14.2.2 Objects

Operations in R operate on objects, sometimes known as variables. Objects are characterized by their names and their contents; with the contents in turn being characterized by attributes specifying the kind of data contained in the object.

The R type system has evolved over time and includes the terms *mode* (a higher level view of the value representation, at least sometimes, than *typeof*, e.g., `integer` and `double` have mode `numeric`), *storage.mode* (a concept going back to the S language) and *typeof* (the underlying representation used by the C implementation of the language).

The `mode`, `storage.mode` and `typeof` functions return a string containing the respective information, e.g., `numeric`, `integer` or `function`.

The length of an object is the number of elements it contains. The `length` function returns the number of elements contained in its argument.

The assignment operator creates an object, with the object name being the left operand and its value and type being that of the right operand.

## 14.3 Operations on vectors

### 14.3.1 Creating a vector/array/matrix

An R vector can be thought of as a one-dimensional array. Vectors are indexed starting at 1 (not zero) and it is possible to add additional elements to a vector, but not remove an existing element.

```
x = 2                      # new vector containing one value
x = c(2, 4, 6, 8, 10)      # new vector containing five values
# new vector containing the contents of x and two values
x = c(x, 12, 14)
y = vector(length=5)        # new vector created by function call
y = 3:8                     # same as c(3, 4, 5, 6, 7, 8)
z = seq(from=3, to=13, by=3) # create a sequence of values
# All elements converted to a common type
z = c(1, 2, "3")           # String has the greater conversion precedence
```

Multidimensional arrays are created using the `array` function, with the common case of 2-dimensional arrays supported by a specific function, i.e., the `matrix` function.

```
> # create 3-dimensional array of 2 by 4 by 6, initialized to 0
> a3=array(0, c(2, 4, 6))
> matrix(c(1, 2, 3, 4, 5, 6), ncol=2) # default, populate in column order
 [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> # specify the number of rows and populate by row order
> matrix(c(1, 2, 3, 4, 5, 6), nrow=2, byrow=TRUE)
 [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
> x = matrix(nrow=2, ncol=4) # create a new matrix
> y = c(1, 2, 3)
> z = as.matrix(y) # convert a vector to a matrix
> str(y)
num [1:3] 1 2 3
> str(z)
num [1:3, 1] 1 2 3
```

### 14.3.2 Indexing

One or more elements or a vector/array/matrix can be accessed using indexing. Accesses to elements that do not exist return NA. Negative index values specify elements that should not be included in the returned value.

The zeroth element returns an empty vector.

```
> x = 10:19
> x[2]
[1] 11
> x[-1]                      # exclude element 1
[1] 11 12 13 14 15 16 17 18 19
> x[12]                       # there is no 12'th element
[1] NA
> x[12]=100                  # there is now
> x
[1] 10 11 12 13 14 15 16 17 18 19 NA 100
```

Multiple elements can be returned by an indexing operation.

```
> x = 20:29
> x[c(2,5)]                  # elements 2 and 5
[1] 21 24
> y = x[x > 25]              # all elements greater than 25
```

```

> y
[1] 26 27 28 29
> # The expression x > 25 returns a vector of boolean values
> i = x > 25
> i
[1] FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE
> # an element of x is returned if the corresponding index is TRUE
> x[i]
[1] 26 27 28 29

```

Matrix indexing differs from vector indexing in that an out-of-bounds access generates an error.

```

> x = matrix(c(1, 2, 3, 4, 5, 6), ncol=2)
> x[2, 1]
[1] 2
> # x[2, 3] need to be able to handle out-of-bounds subscripts in Sweave...
> x[, 1]
[1] 1 2 3
> x[1, ]
[1] 1 4
> x[3, ]=c(0, 9)
> x
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    0    9
> x=cbind(x, c(10, 11, 12))  # add a new column
> x
     [,1] [,2] [,3]
[1,]    1    4   10
[2,]    2    5   11
[3,]    0    9   12
> x=rbind(x, c(5, 10, 20))  # add a new row
> x
     [,1] [,2] [,3]
[1,]    1    4   10
[2,]    2    5   11
[3,]    0    9   12
[4,]    5   10   20

```

### 14.3.3 Lists

The difference between a list and a vector is that different elements in a list may have different modes (types) and existing elements can be removed.

```

> x = list(name="Bill", age=25, developer=TRUE)
> x
$name
[1] "Bill"

$age
[1] 25

$developer
[1] TRUE
> x$name
[1] "Bill"
> x[[2]]
[1] 25
> x = list("Bill", 25, TRUE)
> x
[[1]]
[1] "Bill"

[[2]]
[1] 25

```

```

[[3]]
[1] TRUE
> y = unlist(x) # convert x to a vector, all elements are converted to strings
> y
[1] "Bill" "25"   "TRUE"
> x = list(name="Bill", age=25, developer=TRUE)
> x$sex="M"  # add a new element
> x
$name
[1] "Bill"

$age
[1] 25

$developer
[1] TRUE

$sex
[1] "M"
> x$age = NULL # remove an existing element
> x
$name
[1] "Bill"

$developer
[1] TRUE

$sex
[1] "M"

```

The [[ ]] operator returns a value while [ ] returns a sublist (which has mode list).

#### 14.3.4 Data frames

From the perspective of a programmer coming from another language, it seems obvious to think of a data frame as behaving like a matrix, and in some cases it can be treated in this way (e.g., when all columns have the same appropriate type functions expecting a matrix argument may work). However, a better analogy is to treat it like an indexable structure type (where different members can have different types).

The `read.csv` functions reads a file containing columns of different values and returns a data frame.

When indexing a data frame like a matrix, elements are accessed in row-column order (not the column-row order found in C-like languages). The following code selects all rows for which the `num` column is greater than 2.

```

> x = data.frame(num=c(1, 2, 3, 4), name=c("a", "b", "c", "d"))
> x
  num name
1   1    a
2   2    b
3   3    c
4   4    d
> # Have to remember that rows are indexed first and also specify x twice
> x[x$num > 2, ]
  num name
3   3    c
4   4    d
> # No need to remember row/column order and only specify x once
> subset(x, num > 2)
  num name
3   3    c
4   4    d

```

If one or more columns contain character mode values, `read.csv` will create a factor rather than a vector.

### 14.3.5 Symbolic forms

```
exp = expression(x/(y+z))
eval(expr) # evaluate expression using the current values of x, y and z
```

Uses of expression objects include specifying which vectors in a table to plot in a graph and including equations in graphs, for instance:

```
text(x, y, expression(p == over(1, 1+e^(alpha*x+beta))))
```

will cause the following equation to be displayed at the point (x, y).

$$p = \frac{1}{1 + e^{\alpha x + \beta}}$$

The D function takes an expression as its first argument and based on the second argument returns its derivative:

```
> D(expression(x/(y + z)^2), "z")
-(x * (2 * (y + z)) / ((y + z)^2)^2)
```

### 14.3.6 Factors and levels

Statisticians found that when manipulating non-numeric values (e.g., names) it can be convenient to map them to integer values and manipulate these integers. In programming terminology a variable used to represent one or more of these integer values could be said to have a *factor* type, with the actual numeric values known as *levels* (a parallel can be drawn with the enumeration types found in C++ and C, except these assign names to integer values).

```
> factor(c("win", "win", "lose", "win", "lose", "lose"))
[1] win win lose win lose lose
Levels: lose win
```

Some operations implicitly convert a sequence of values to a factor. For instance, `read.csv` will, by default, convert any column of string values to a factor; this conversion is a simplistic form of hashing, and (when a megabyte was considered a lot of memory) was once driven by the rationale to save storage space. These days the R implementation uses more sophisticated hashing and we have to live with the consequences of historical baggage.

Operations of objects holding values represented as factors sometimes has surprising effects for those unaware of how things used to be.

## 14.4 Operators

Operators in R follow the same precedence rules as Fortran, which in some cases differ from the C precedence rules (which most commonly languages now use). An example of this difference is: `!x ==y` which is equivalent to `!(x ==y)` in R, but equivalent to `(!x) ==y` in C-like languages (if x and y have type boolean, there is no effective difference, but expressions such as: `!1 ==2` produce a different result).

A list of operators and their precedence can be obtained by typing `?Syntax` at the R command line.

Within an expression operand evaluation is left to right, except assignment which evaluates the right operand and then the left.

In most cases, all elements of a vector are operated on by operators:

```
> c(5, 6) + 1
[1] 6 7
> c(1, 2) + c(3, 4)
[1] 4 6
> c(7, 8, 9, 10) + c(11, 12)
[1] 18 20 20 22
> c(0, 1) < c(1, 0)
[1] TRUE FALSE
```

in the last two examples *recycling* occurs, that is the elements of the shorter vector are reused until all the elements of the longer vector have been operated on.

The operators **&&** and **||** differ from **&** and **|** in that they operate on just the first element of their operands and return a vector containing one element, e.g., `c(0,1) && c(1,1)` returns the vector FALSE.

The base system includes a set of `rfunc[bitw??]` functions that perform bitwise operations on their integer arguments and there is the `bitops` package.

Operators	Description
<code>:: :::</code>	access variables (right operand) in a name space (left operand)
<code>\$ @</code>	component / slot extraction (member selection has lower precedence than subscripting in C-influenced languages)
<code>[ [[</code>	array and list indexing
<code>^</code>	exponentiation (associates right to left)
<code>- +</code>	unary minus and plus
<code>:</code>	sequence operator
<code>%any%</code>	special operators (%% and %/% has the same precedence as * and / in C-influenced languages)
<code>* /</code>	multiply and divide
<code>+ -</code>	(binary) add subtract
<code>&lt; &gt; &lt;= &gt;= == !=</code>	relational and equality (non-associative; equality has lower precedence in C-influenced languages)
<code>!</code>	negation (greater precedence than any binary operator in C-influenced languages)
<b>&amp; &amp;&amp;</b>	and of all and first
<b>  </b>	or of all and first
<code>~</code>	as in formulae
<code>-&gt; -&gt;</code>	local and global rightwards assignment
<code>=</code>	assignment (associates right to left)
<code>&lt;- &lt;-</code>	local and global assignment (associates right to left)
<code>?</code>	help (unary and binary)

Table 14.1: Operators listed in precedence order.

The character used for exclusive-or in C-influenced languages, `^`, is used for exponentiation in R; the `xor` function performs an exclusive-or of its operands.

The `[` and `[[` operators differ by more than being array and list indexing. The result of the index `x[1]` has the same type as `x` (i.e., the operation preserves the type), while the result of `x[[1]]` is a simplified version of the type of `x` (if simplification is possible).<sup>i</sup>

```

> x = c(a = 1, b = 2)
> x[1]
a
1
> x[[1]]
[1] 1
> x = list(a = 1, b = 2)
> str(x[[1]])
List of 1
  $ a: num 1
> str(x[[1]])
num 1
> x = matrix(1:4, nrow = 2)
> x[1, ]
[1] 1 3
> x[1, , drop = FALSE]
     [,1] [,2]
[1,]    1    3
> # x[[1, ]] is not allowed
>
> df = data.frame(a = 1:2, b = 1:2)
> str(df[1])

```

<sup>i</sup> Out of bounds handling is also different, but I'm sure readers' don't do that sort of thing.

```
'data.frame': 2 obs. of 1 variable:
$ a: int 1 2
> str(df[[1]])
int [1:2] 1 2
> str(df[, "a", drop = FALSE])
'data.frame': 2 obs. of 1 variable:
$ a: int 1 2
> str(df[, "a"])
int [1:2] 1 2
```

### 14.4.1 Testing for equality

In addition to the equality operators the base system includes two equality related functions, `identical` and `all.equal`.

```
> x = 1:5 ; y = 1:5
> x == y # Return the result of equality test for each element
[1] TRUE TRUE TRUE TRUE TRUE
> identical(x, y) # Return a single value denoting exact equality
[1] TRUE
> 1L == 1 # 1L is stored internally as an integer, 1 is stored as a double
[1] TRUE
> identical(1L, 1) # identical requires the stored type be the same
[1] FALSE
> 0.9 == (1.1 - 0.2) # could be affected by lack of precision
[1] FALSE
> all.equal(0.9, 1.1 - 0.2) # do a fuzzy compare
[1] TRUE
> all.equal(0.9, 1.1 - 0.2, tolerance=0) # find out much fuzz there is
[1] "Mean relative difference: 1.233581e-16"
```

The default tolerance used by the `all.equal` function is `.Machine$double.eps ^ 0.5`.

Comparisons against NA always returns NA and the `is.na` function has to be used to check for this quantity; the `anyNA` function returns TRUE/FALSE if its argument contains an NA.

### 14.4.2 Assignment

The following are four ways of the ways of assigning a value to a variable in R:

```
x <- 3 # Operator used by people who follow the herd
x <<- 3 # Assigns to the x at global scope

3 -> x # Rarely encountered outside descriptions of the language

x = 3 # Supported since R version 1.4
```

Many R books and articles use the token `<-`. Developers are used to seeing the `=` token and with nothing other than conformity to existing R usage to recommend the alternative, the token that developers are already very familiar with, is used in this book.

There is one context where `=` does not behave like normal assignment. R supports the use of parameter names in arguments to function calls to explicitly specify that a named parameter is to be assigned a given value. In the context of a function argument list the left operand of `=` is treated as the name of a parameter and the right operand as the value to be assigned. An error is flagged if the function definition does not have a parameter having the specified name.

```
func = function (a, b, c) a + b * c

func(2, 3, 9)
func(c=9, b=3, a=2)

func(d=3, 4, 5) # no parameter named d, an error is raised

# use <- if the intent is to assign to d and pass this value as an argument
func(d<-3, 4, 5)
```

## 14.5 The R type (mode) system

R supports values having the following basic types (R also has the concept of *mode*, which is based on semantics rather than underlying representation, e.g., the mode function returns numeric where typeof returns either integer or double):

- NULL:
- raw: essentially uninterpreted byte values,
- logical: holds one of the values: TRUE, FALSE, T or F. The conversion `as.logical(any_n_on_zero_value)` returns TRUE,
- character: what many other languages call a string type,
- integer: the only integer type uses 32 bits (NA is represented using the most negative value, so this value is not available as an integer; trying to generate this or any other value outside of the representable value of a 32-bit integer will result in a value having a real type),
- double: the only floating-point type contains 64 bits,
- complex: contains a real and imaginary double type,

An object may be reported to have one of these basic types, but it may actually be a vector or array of this type.

More complicated types may be created, such as lists, data frames, etc.

### 14.5.1 Converting the type (mode) of a value

It is often possible to convert the mode (type) of a value by calling the `as.some_mode` function, where `some_mode` is the name of a mode, e.g., `integer`. If a conversion fails, NA is returned.

#### Conversion precedence

```
NULL < raw < logical < integer < real < complex < character < list < expression
```

## 14.6 Statements

R contains the usual language constructs that look like statements, but can behave like expressions:

- **function**: defines a function, whose value has to be assigned to an object:  
`f=function(p1, p2) {return(p1+p2)}`
- blocks of code are bracketed using the punctuation pair: { and },
- ; (semicolon) is required to delimit multiple expressions on the same line, but is otherwise optional,
- **if**: which takes an optional **else** arm (there no **then** keyword, but there is an **ifthenelse** function),
- **for**: which has the form `for (i in x)`, where x is a vector (such as `1:10`),
- **while** and **repeat** loops are available,
- loops may be terminated using **break** keyword or the break function, and may be continued at the next iteration using the **next** keyword or next function,
- **return** is a function: `return(1+return(1))` returns the value 1,
- **switch** is a function.

## 14.7 Defining a function

```
> g=1 # a global variable
> f = function(p1, p2) # define a function and assign it to f
+ {
+ l=g # Value access, check lexical and dynamic scope for g
+ g=2 # Assignment: only check local scope, if no variable exists, create one
+
+ m=h # h is dynamically in scope
+
+ return(return(1)+1) # return is a function call
+ }
> h=2 # another global variable
> f(1, 2)
[1] 1
> g
[1] 1
> h=3 # At global scope, so must be global variable
```

Argument evaluation is lazy, that is they are evaluated the first time their value is required.

The ... token specifies that a variable number of unknown arguments may be passed.

```
unk_args=function(...)
{
a=list(...) # Convert any arguments passed to a list of values
# Access the list of values in a
}
```

## 14.8 Commonly used functions

Technically every operation is a function call (so '+'(1, 2) and 1+2 are equivalent), but not all function calls have equivalent operator tokens.

```
> x = 1:10
> if (any(x > 7)) print("At least one value greater than 7")
[1] "At least one value greater than 7"
> if (all(x > 0)) print("All values greater than zero")
[1] "All values greater than zero"
> rep(1:2, 3)
[1] 1 2 1 2 1 2
```

`head/tail` mimics the behavior of the unix `head/tail` program,

`length` returns the number of elements in its vector argument,

`nrow/ncol` return the number of rows/columns in the data frame argument (NROW/NCOL gracefully handle vector arguments),

`order` returns a vector containing an index into the argument in the order needed to sort the argument values,

`str` lists the columns in a variable, along with their type and the first few values in each row; it provides a quick way of verifying that columns have the expected type.

`which` returns a vector of values containing the index of the argument values that are true,

`methods`: list functions overloaded on the argument name

`installed.packages`: list all installed packages

`getwd, setwd`:

`list.files, list.dirs`:

`ls` lists variables that exist in the current environment,

`system.time, proc.time`:

## 14.9 Input/Output

Functions are available for reading data having a variety of formats (e.g., comma separated values) from all the common data sources (e.g., files, databases, web pages). In some cases the contents of a compressed file will be automatically uncompressed before reading. The data read is often returned as a single object.

Many functions try to automatically deduce the datatype of the data read, e.g., whether it is integer, real, sequence of characters, etc. Sometimes the datatype selected is not correct and work has to be done to ensure the data is treated as having the desired type; `read.csv` bases its decision on the type of the columns by analysing the first 6 or so lines of the file.

Some functions in the base system, e.g., `read.csv`, convert columns containing string values to factors; the original intent was presumably to reduce the storage needed to hold the data. A column of factors does not always behave the same as a column of strings and this default conversion behavior is now a liability. Passing the argument `as.is=TRUE` stops values being converted to factors (it is used in all the example code).

```
data=read.csv("measurements.csv.gz", as.is=TRUE) # file will be uncompressed

data=read.csv("measurements.csv", sep="|", as.is=TRUE) # change separator

data=read.csv("https://github.com/Derek-Jones/ESEUR-code-data/blob/master/benchmark/MST")
```

The first line of the input file is assumed to denote the column names, and `header=FALSE` has to be specified to switch off this behavior.

All characters on an input line after, and including, the comment character, `#`, are ignored (various options interact with this behavior, including the `comment.char` option which can be used to change the character used).

If data is not already in a form that can be easily processed by R, it may be simpler to convert it by using a language or tool that you are already familiar with, rather than using R.

Output is supported by a wide range of functions: `print` performs relatively simple formatted output (the `format` function can be used to create more sophisticated formatting, that can then be output), the `cat` function performs relatively little formatting but is more flexible and in particular does not terminate its output with a newline, the `sink` function can be used to specify an alternative location to write console output, and there are often `write` equivalents of the `read` functions such as `write.csv`.

The R environment includes a simple spreadsheet like editor for manual data entry and patching existing data.

```
scores = edit(scores) # use built in spread-sheet like editor
```

### 14.9.1 Graphical output

There may be more functions supporting graphical output in R than textual output. Perhaps the most commonly used graphical output function is `plot`. This function often does a surprisingly good job of producing a reasonable graphical representation of the data. Overloaded versions of this function are sometimes provided by packages, to plot data having a particular class created by the package.

By default, graphical output is sent to the console device; this behavior can be overridden to produce a file having a particular format, e.g., `pdf`, `jpeg`, `png` and `pictex`. The list of supported output devices varies across the operating systems on which R runs.

The behavior of the `plot` function can be influenced by previous calls to the `par` function, which set configurable options.

Various packages proving graphical output are available, with the `ggplot` package probably being the most commonly used by frequent R users.

## 14.10 Other uses for R

While the target of R's domain specialised functionality is statistical data analysis, there are other application domains where this functionality could be useful (but do not warrant effort needed to learn R).

A variety of functions designed for manipulating the rows and columns of delimited data files are available; see [Rlang/Top500.R](#).

A useful technique for spotting whether a file contains compressed data, e.g., a virus hidden in a script by compressing it to look like a jumble of numbers. Compressed data contains a uniform distribution of byte values (after all, compression is achieved by reducing apparent information content), your mileage may vary between compression techniques.

The following code reads a complete file, applies a sliding window to the data and then plots it.

```
window_width=256 # if less than 256, divisor has to change in plot call

plot_unique=function(filename)
{
  t=readBin(filename, what="raw", n=1e7)

  # Sliding the window over every point is too much overhead
  cnt_points=seq(1, length(t)-window_width, 5)

  u=sapply(cnt_points, function(X) length(unique(t[X:(X+window_width)])))
  plot(u/256, type="l", xlab="Offset", ylab="Fraction Unique", las=1)

  return(u)
}

dummy=plot_unique("http://www.coding-guidelines.com/R_code/requirements.tgz")
```

## 14.11 Very large datasets

While most existing software engineering datasets tend to be relatively small, exceptions may occur from time to time. A variety of techniques are available for handling large datasets, including the following:

- the `bigmemory` package provides software defined memory management (i.e., swapping data between memory and main storage). The `bigtabulate` package and other `big` packages contains functions that perform commonly used operate on this data.
- the `data.table` package extends `data.frames` to support up to 100G of storage,

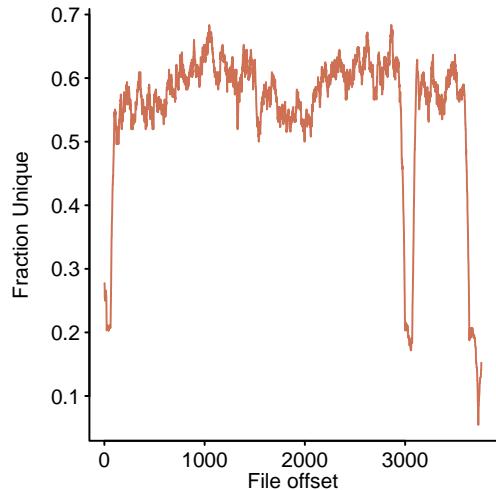


Figure 14.2: The unique bytes per window (256 bytes wide) of a pdf file. [code](#)

## 14.12 Debugging R code

The `traceback` function...

Packages to help debug R programs include: `RUnit` for unit testing, `covr` for code coverage,...



# Chapter 15

## Data preparation

### 15.1 Introduction

The most important question to keep asking yourself while examining, preparing and analyzing any data is: Do I believe this data?

Do patterns appear when none are expected, are expected patterns absent, are human errors missing from the raw data, does the data collector believe whatever they are told, does the measurement process create incentives for people to game it?

Books and presentations on data analysis rarely mention that a significant amount of effort often has to be invested in preparation data (perhaps 80% or more of analysis effort), getting it into a form that can be input to the chosen statistical analysis techniques.<sup>i</sup>

Perhaps the largest task within data preparation is data cleaning, an often overlooked<sup>845</sup> aspect of data analysis that is an essential part of the workflow needed to avoid falling foul of the adage: Garbage in, garbage out.

Domain knowledge is essential for data cleaning; values have to be understood in the context in which they occur. The fact that many data cleaning activities are generic does not detract from the importance of domain knowledge. For instance, software knowledge tells us that 1.1 is not a sensible measurement for lines of code (in the NASA MDP dataset), talking to developers at a company that they don't work at weekends (e.g., the dates in the 7digital data) and knowing that system support staff used the Unix `pwd` command to check that the system was operational (an analysis of job characteristics for a NASA supercomputer<sup>418</sup> has to first remove 56.8% of all logged jobs).

Data cleaning is often talked about as-if it is something that happens before data analysis, in practice, the two activities intermingle; the time spent checking and cleaning the data provides insights that lead to a better understanding of the kinds of analysis that might be applicable, also, results from a preliminary analysis can highlight data needing to be cleaned in some way. Data preparation is discussed here in its own chapter, as-if it was performed as a stand-alone activity, in order to simplify the discussion of the material in other chapters.

Once cleaned the data may need to be restructured, e.g., rows/columns contained in different files merged into a single data table, or the row/columns in an existing table reorganised in some way. The required structure of the data is driven by the operations that will be performed on it (e.g., finding the median value of some attribute) or the requirements of the library function used to perform the analysis.

It may be necessary to remove confidential information from the data or to remove information that might be used to identify individuals. Datasets do not exist in isolation and it may be possible to combine apparently anonymous datasets to reveal information;<sup>ii</sup>  $k$ -anonymity and  $l$ -diversity are popular techniques for handling this situation (in a  $k$ -anonymized dataset each record is indistinguishable from at least  $k - 1$  other records and  $l$ -diversity requires at least  $l$  distinct values for each sensitive attribute). Techniques for anonymizing data are not

<sup>i</sup> This chapter appeared immediately after the Introduction in early versions of this book, but in response to customer demand was moved here (the last chapter); people want to read about the glamorous stuff, data analysis, not the grunt work, data cleaning.

<sup>ii</sup> 63% of the US population can be uniquely identified using only gender, ZIP code and full date of birth.<sup>513</sup>

covered here; Fung et al<sup>463</sup> survey techniques for privacy-preserving data publishing, Templ et al<sup>1343</sup> provide an introduction to statistical disclosure control and the `sdcMicro` package.

While a lot of software engineering data comes from measurements made by programs, some is still sourced from human written records (which can contain a substantial number of small mistakes<sup>693</sup>).

Data cleaning involves a lot of grunt work that often requires making messy trade-offs and having to make do. Tools are available to reduce the amount of manual work involved, but these may require placing some trust in the tool doing the right thing. The following are a few of the available tools:

- OpenRefine<sup>1036</sup> (was Google Refine) reads data into a spreadsheet-like form and supports sophisticated search/replace and data transformations editing,
- the `editrules` package checks that values are consistent with user specified consistency rules involving named columns (e.g., `total_fruit == total_apples+total_oranges`),
- the `deducorrect` package performs automatic value transformation based on user specified consistency rules about the column values that must be met (e.g., failure to meet the condition `total_x > 0` will result in any value in that column having a negative sign removed); this package can also impute missing values,
- there are a variety of special purpose packages that handle domain specific data, e.g., the `CopyDetect` package detects copying of exam answers in multi-choice questions.

?

### 15.1.1 Document cleaning operations

The changes made to the original data during the cleaning process need to be documented; this change log serves a variety of purposes, including:

- enabling third-parties to check that the changes are reasonable and don't substantially alter the results of the analysis,
- enabling a potential source of uncertainty to be checked when multiple outlets publish results based on the same dataset, i.e., if there are differences in the results it is possible to check that these differences are not primarily the result of differences in cleaning operations,
- providing confidence to users of the final results of the analysis that the researcher doing the work is competent, i.e., that cleaning was performed.

Ideally the operations performed on the original data to transform it into what is considered a clean state are collected together as a script for ease of replication.

Some cleaning activities are trivial and yet need to be performed to prevent the analysis being overwhelmed by what appears to be lots of special cases. For instance, an analysis<sup>55</sup> of different company's response to vulnerabilities reported in their projects, started with data that sometimes used slightly different ways of naming the same company. The company names data had to be cleaned to ensure that one name was consistently used to denote each organization (see `data-check/patch-behav.R`).

The publicly available NASA Metrics Data Program (MDP) dataset contains fault data on 13 projects and has been widely used in academic research (a literature survey for the period 2000 to 2010<sup>573</sup> found that 58 out of 208 papers used it). This dataset contains many problems<sup>541</sup> that need to be sorted out, e.g., columns with all entries having the same value (suggesting a measurement or conversion error has occurred), duplicate rows, missing values (many occurred in rows calculated from other rows and involved a divide by zero), inconsistent values (e.g., number of function calls being greater than the number of operators) and nonsensical values (e.g., lines of code having fractional values).

Despite the extensive work needed to clean the MDP dataset to get it into a reliable state, the authors of many of the published papers using this dataset have either not cleaned it or have only given a cursory summary of their cleaning activities<sup>153</sup> (“ . . . removes duplicate

tuples . . . along with tuples that have questionable values . . . ”, does not specify what values were questionable). Consequently, even although the original dataset is publicly available it is difficult to compare the results published in different papers because no information is available on what, if any, data cleaning operations were performed; so much of the data is in need of cleaning that any results based on an uncleaned version of this dataset must be treated with suspicion.

See [data-check/NASA.MDP-data\\_check.R](#) for examples of various integrity checks performed on the MDP dataset.

A survey of 682,000 unique Android devices in use during 2015, by OpenSignal,<sup>1037</sup> included the screen height and width reported by the device (upper plot in Figure 15.1). Many devices appear to have greater width than height, particularly those with smaller screens. Perhaps the device owners are viewing the OpenSignal website with their phones in landscape mode (lower plot in Figure 15.1 switches the dimensions so that height has the larger value; switched values in red).

Like their source code, the fault repositories of open source projects are publicly available and the repositories of larger projects are a frequent source of data for fault analysis/prediction researchers.

A study by Herzig, Just and Zeller<sup>610</sup> manually classified over 7,000 issue reports extracted from the fault repositories of seven large Java projects. They found that on average 42.6% of reports had been misclassified, with 39% of files marked as defective not actually containing any reported fault (the implication being that any fault prediction models built using this uncleaned data are likely to be misleading at best and possibly very wrong). An earlier analysis<sup>611</sup> had found that between 6 and 15% of bug fixing changes addressed more than one issue.

Possible reasons for the misclassification include: the status of an issue not being specified when the initial report is filed, resulting in the default setting of *Bug* being used; issue submitters having the opinion that a missing feature is a bug (request for enhancement was the most commonly reclassified status); and bug reporting systems only supporting a limited number of different issue statuses (forcing the submitter to use an inappropriate status).

This study shows, at least for the fault repositories of seven large Java projects, that data cleaning is essential for any analysis based on this data to be reliable. The study also highlighted how much effort data cleaning consumes; the work was performed independently by two people and took a total of 725 hours (90 working days).

Sometimes the measured values from one or more subjects (e.g., people or programs) are remarkably different from the measured values of other subjects. It can be tempting to clean the data by removing the measured value for these subjects from the dataset. A study by Müller and Höfer<sup>978</sup> removed data on seven out of 18 subjects because they considered the performance of these subjects was so poor that they constituted a threat to the validity of the experiment (whose purpose was to compare the performance of students and professional developers). This kind of activity might be classified as outlier removal or manipulating data to obtain a desired result, either way, documenting the cleaning activity allows the audience to decide.

## 15.2 Outliers

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”.<sup>588</sup>

Without some knowledge of the mechanism that produced a given sample it is not possible to label any observation as suspicious, although it might be possible to claim that the observations were generated by more than one mechanism. If the percentage of deviate observations is small, the mechanism that produced them might be labelled as unimportant and those observations excluded from the subsequent analysis; either way, document the situation and let your audience decide.

In some applications the deviate observations are the ones of interest,<sup>230</sup> e.g., intrusion detection and credit card fraud. This subsection covers the case where deviate observations are unwanted; some later subsections cover software engineering situations where deviate observations are themselves the subject of study.

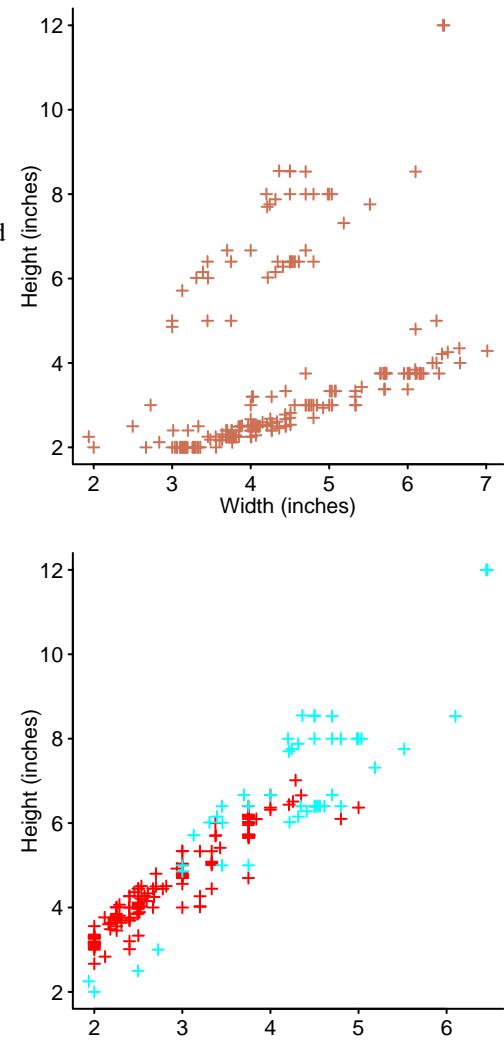


Figure 15.1: Screen height and width reported by 682,000 unique devices that downloaded an App from OpenSignal in 2015 (upper), reported measurements ordered so height always the larger value (lower). Data from OpenSignal.<sup>1037</sup> [code](#)

Two commonly used ways of handling outliers are:

- using a statistical technique that does not assign too much weight to observations that deviate from the common pattern followed by other observations. Techniques capable of performing the desired statistical analysis are not always available, but when R functions implementing them are available they are discussed in the appropriate section,
- detecting and excluding outliers from the subsequent analysis. Traditionally outliers have been manually selected and excluded from subsequent analysis. This approach can work well when the sample contains a small amount of data and the person doing the detection has sufficient domain knowledge. There are a variety of functions that automate the process of outlier selection and handling, some of these are discussed below.

Another definition of outlier detection<sup>230</sup> is “ . . . the problem of finding patterns in data that do not conform to the expected normal behavior.” This definition requires that an expected normal behavior be known, along with a method of comparing values for *outlyingness*.

Figure 15.2 shows a suspicious spike in the number of daily reported vulnerabilities recorded in the US National Vulnerability Database for 2003. Perhaps all vulnerabilities that had been reported, but not yet fully processed, were simply published for the public to see at the end of the year? In this sample the analysis is highlighting a subset of the data reported on one day as a suspected outlier.

The date when an event occurred may appear unlikely, based on domain knowledge, e.g., staff rarely work at weekends. The following is a count of the number of features recorded as being Done, in a company using an Agile process,<sup>1162</sup> for each day of the week. Monday is day 0 and the counts for Saturday/Sunday should be zero; the non-zero values suggest a 2-4% error rate, comparable with human error rates for low stress/non-critical work. [code](#)

```
> table(Done_day %% 7)
  0   1   2   3   4   5   6 
 670 708 669 716 447 12  16
```

Should outliers be removed from a sample?

While removing outliers may improve the quality of the fit to an equation, does it improve the quality of the fit of the model to reality?

Without understanding the process that generated the data there is no justification for removing any value.

The real issue with outliers is the impact they have on the final result. In a large sample a few unusual values are unlikely to have any real impact data.

However, outliers are handled, the decision needs to be documented, giving either the reason for excluding observations from the analysis or the reason for keeping them and their impact on the final result.

## 15.3 Malformed file contents

The first sign that data or its organization in a file is malformed in some way, usually occurs when the variable into which the file contents have been read does not have the expected contents (e.g., incorrect number of columns or the type of data in one or more columns is not as expected, such as strings where a number should have occurred).

The `str` function provides a quick way of verifying that columns in a `data.frame` have the expected type. For instance, unexpected characters in the input data can result in some column values being treated as strings when a numeric type is expected.

File contents formatting issues to watch out for include:

- functions for reading data in R (e.g., `read.csv` and `read.table`) often use the first few lines of file being read as the format to when reading the rest of the file, i.e., the number of columns contained in each row and the datatype of the values in each column. If there are one or more rows that do not follow the format selected at the start of the file (e.g., different number of column delimiters, perhaps the result of non-delimited strings such as a missing pair of quote characters) then subsequent values may not appear in the correct column or be converted to a different type,

- termination delimiter missing from a string value, which can result in the following row being treated as part of the current row (because the newline is treated as part of the string),
- cut-and-pasting of data introducing conversion errors, e.g., the digit zero treated as the letter D or G.

A variety of ad-hoc techniques are available for locating the cause of problems. For instance, the following code will convert all values that do not have the format of a number to NA, which are easily located using base-library function support for processing NAs, and their row index listed using `which`:

```
which(is.na(as.number(as.character(data.frame$column_name))))
```

The `complete.cases` function returns a vector specifying which of the rows in its `data.frame` argument are complete, i.e., do not contain any NA values; the `na.omit` function returns a copy of its argument with any rows containing NA omitted.

## 15.4 Missing data

Missing data (for instance, a survey where the entry for a person's age is empty) is often the rule rather than the exception and whole books have been written on the subject. Missing data may be disguised in that it appears as a reasonable looking value,<sup>1069</sup> e.g., zero when the range of possible legitimate values includes zero.

The starting point for handling missing data is to normalise how it is denoted to the representation used by R, i.e., NA (Not Available). Normalisation ensures that all missing values are treated consistently; special case handling of NA is built into R and many functions include options for handling NA.

A wide variety of different representations for missingness are likely to be encountered (e.g., special values that cannot occur as legitimate data values such as 9999, "#N/A", "missing" or no value appearing between two commas in a comma separated list), and it is not uncommon for different columns within a dataset to use different representations (because they originate from different sources of measurement).

Missing values are sometimes *disguised* in that they look reasonable, e.g., a value of zero (where domain knowledge does not rule out zero as a realistic measurement value) or when data input only provides a two item choice, such as male/female, with one item being the default and thus appearing as the missing value when no explicit choice is made. Disguised missing values can have a significant impact on the results.<sup>1069</sup>

The following shows one way of changing a known representation of missing value to NA (the second form would be necessary if 9999 could appear as a legitimate value in a column other than `size`):

```
data[ data == 9999 ] = NA # set all elements having value 999 to NA

# set all elements of column size having value 999 to NA
data$size[ data$size == 9999 ] = NA
```

Once missing values have been explicitly identified it is possible to move on to deciding whether to ignore these cases or to replace NA with some numeric value. Some algorithms can handle missing values while others cannot; R functions vary in their ability to handle missing values. A few techniques for selecting the replacement value are discussed below.

The R base I/O functions, such as `read.csv`, have conventions for handling the case of zero characters appearing between the delimiters on each line of a file. The behavior depends on what type the values in a particular column are assumed to have. For columns considered to have a numeric type the zero character case is treated as if NA occurred between the delimiters, while for columns considered to have a string type the zero character case is treated as the empty string rather than NA (i.e., treated the same as the string ""). These functions support a variety of options for changing the default handling of leading/trailing white-space between delimiters and the handling of zero characters.

For instance, reading a file containing the columns appearing below left has the same effect as reading a file containing the columns appearing below right:

X,Y_str,Z	X,Y_str,Z
1,"abc",2.2	1,"abc",2.2
2,,3.1	2,"",3.1
,NA,2	NA,NA,2

The `table` function is commonly used to count occurrences of values, but by default does not include NA in its count; the `useNA` needs to be used to explicitly specify that they be included:

```
table(data$some_column, useNA="ifany") # limit the count to one column
```

This one column use can be expanded to cover every column in `data`. If the output is too voluminous, the number of columns processed will need to be reduced or the call to `table` replaced by a call to `tabulate` which provides more options to control behavior:

```
sapply(colnames(data), function(x) table(data[, x], useNA="ifany"))
```

While there may be documentation specifying how missing values are represented, such details are often not written down. An analysis of a dataset using the above code may show a suspiciously large number of values such as 9999 or -1 (for an attribute that can never be negative), something that warrants further investigation.

### 15.4.1 Handling missing values

When deciding what to do about missing values it is important to try to understand why the values are missing. The following categories are commonly encountered in the analysis of missing data:

- Missing completely at random (MCAR): As the name suggests, the selection of missing values occurred completely at random. Statistically this is the most desirable kind of missingness because it means there is no bias in the missing values.
- Missing at random (MAR): This sounds exactly like MCAR, but it is not completely random in the sense that the choice of which values are missing is influenced by other values in the sample. For instance, the level of seniority may correlate with the likelihood that survey questions about salary are answered,
- Missing not at random (MNAR): This missingness could be as random as MAR, with the one difference being that the choice of missing values is influenced by values not in the sample. For instance, the name of the developer who originally wrote the code referenced in a fault report may be missing if that developer is a friend of the person reporting the fault, with friendship not being a recorded in the sample.

The following code can be used to get a rough estimate of the correlation between the rows of a `data.frame` that contain missing values (Figure 8.7 illustrates a way of visualizing this information):

```
x=is.na(some_data_frame)
# highlight rows having some, but not all, missing values
cor(subset(x, sd(x) > 0))
```

Many analysis techniques handle missing values by ignoring the rows or columns that contain them; if the sample has many rows and a low percentage of missing values, then this behavior may not be a problem. However, if the sample contains a large percentage of missing values, any analysis will either have to make do with a smaller number of measurements, be limited to using techniques that can gracefully adapt to missing data (i.e., don't ignore rows containing one or more missing values) or be forced to used estimated values for missing data.

The ideal approach is to use an algorithm capable of handling samples that include missing data.

If missing data is to be replaced by values calculated from values that are present, other data cleaning operations should be performed first to ensure that substitute values are calculated from what are considered to be acceptable values. The issues involved in replacement values are discussed here to concentrate the discussion of this issue in one place.

A quick and dirty method that can be surprisingly effective is to replace missing values by the mean of the values in the corresponding column containing each missing value; alternatively if the data is ordered in some way, the last value appearing before the missing value might be used.

A more sophisticated approach to imputing values is to fill the missing value entries from other values in the dataset. *Amelia* and *VIM* are two packages that provide a variety of functions for visualizing datasets containing missing values and imputing values for these entries (the *VIMGUI* package provides a GUI to *VIM*).

The `aggr` and `marginplot` functions...

As Figure 11.65 shows, data that is missing over a complete range of values, rather than spread over the sample, can have a dramatic impact on fitted models.

When a single object attribute is counted, e.g., number of times the functions in a particular library are called, there is no missing data because there are no other measured attributes that can be missing. However, data may be missing because the sample may not contain any instances of particular cases (which would be seen in a larger sample). Good-Turing smoothing<sup>471</sup> is a technique for adding non-zero counts for unseen items and is a technique that is usually part of an approach to handling missing items, rather than something to use standalone.

## 15.4.2 NA handling by library functions

R functions vary in their ability to handle data.frames containing NA and exhibit the following behaviors:

- behaving in unpredictable ways when NA is encountered,
  - behaving in predictable, but perhaps surprising to the unknowledgeable, ways, e.g., the value of `NA == NA` is `NA` as is `NA != NA`,
- functions that operate on complete rows or columns have a variety of behaviors when they encounter one or more NAs, including:
- supporting a parameter, often called `na.rm`, which can be used to specify that NA should be taken into account/ignored,
  - ignoring rows containing one or more NA, e.g., `glm` ignores these rows by default, but this behavior can be changed using the `na.action` option,
  - making use of information present in rows containing one or more NA, e.g., the `rpart` function,

Some regression model building functions return information associated with individual data points, such as residuals. If the function removes rows containing any NA before building the regression model, then the number of data rows included in the returned model may be less than originally passed in, unless rows containing NA is reinserted (e.g., by using the `naresid` function).

## 15.5 Restructuring data

When the data of interest is contained in several files it may be necessary to read two or more files and merge their contents into a single data.frame.

If two datasets contain share columns (i.e., column names, column ordering and contents are the same) the `rbind` function joins rows together, returning a single data.frame; the `cbind` function performs the same operation for columns.

When two data.frames share common columns the `merge` function can be used to join the two data.frames based on one or more shared column names; there are a variety of options for selecting how merging is performed.

### 15.5.1 Reorganizing rows/columns

The organization of rows and columns in a data.frame may not be appropriate for that used by the library functions used to perform the analysis.

The values in a dataset are sometimes held in a wide format (i.e., a few rows and many columns) and a long format (i.e., many rows and a few columns) is required, or vice versa.

An example of wide format data is that used in Figure 2.4; the IQ test scores have in the following form:

```
test,gender,1,2,3,4,5,6,7,8,9
verbal,Boy,8455,14171,17596,29308,30490,27544,16037,9857,4635
verbal,Girl,5448,10570,15312,28591,32385,30830,18557,11443,5321
quantitative,Boy,3138,19634,18258,29037,23255,30376,16504,12565,5095
quantitative,Girl,2313,16905,19002,32707,26438,32413,15215,10007,3406
non-verbal,Boy,1390,18144,20713,29245,25720,27077,18095,11369,6077
non-verbal,Girl,1165,14370,18564,30488,29342,30458,18387,10450,5075
CAT3,Boy,2505,14505,19556,29917,29607,30327,17960,9392,2787
CAT3,Girl,1813,10927,17872,31059,32867,33269,18016,9041,2394
```

The `melt` function in the `reshape2` package transforms data.frames to a long format, such as the following (only the first 11 lines are shown):

	test	gender	stanine	count
1	verbal	Boy	X1	8455
2	verbal	Girl	X1	5448
3	quantitative	Boy	X1	3138
4	quantitative	Girl	X1	2313
5	non-verbal	Boy	X1	1390
6	non-verbal	Girl	X1	1165
7	CAT3	Boy	X1	2505
8	CAT3	Girl	X1	1813
9	verbal	Boy	X2	14171
10	verbal	Girl	X2	10570
11	quantitative	Boy	X2	19634

which was reorganized using the call (where `b_g_IQ` contains the data):

```
b_g=melt(b_g_IQ, id.vars=c("test", "gender"),
          variable.name="stanine", value.name="count")
```

It is also possible to convert from long to wide format.

## 15.6 Miscellaneous issues

### 15.6.1 Application specific cleaning

The analysis of some kinds of data has acquired established preprocessing procedures; the data is not wrong, but transforming it in some way improves the quality of the subsequent analysis. For instance, before analyzing text, common low interest words (such as “the” and “of”, known as *stop words*) are removed; also words may be stemmed<sup>154</sup> (a process that removes suffixes with the intent of uncovering the root word, e.g., kicked and kicking become kick).

?

### 15.6.2 Different name, same meaning

Typos in character based data may be easy to detect because of constraints on what can appear in any sequence (e.g., the spelling of words). Harder to detect problems include different people using different terminology for the same concept or the same terminology for different concepts.

The SPEC 2006 benchmark results often include a description of the characteristics of the memory used by the computer under test. For historical marketing reasons two scales are commonly used to specify memory performance; the DDR scale is based on peak bandwidth while the PC scale uses clock rate. The SPEC result descriptions are not consistent in their choice of scale and so before any analysis can be performed the values need to be converted to either all DDR form or all PC form. Also, for marketing reasons the numbers have been rounded to reduce the number of non-zero digits; any analysis interested in high accuracy would map all the *marketing* values to their actual values (see [benchmark/scripts/SPEC-memory.awk](#)).

An email address is sometimes the only unique identifying information available, e.g., the list of developers who have contributed to an open source project. The same person may have used more than one email address over the period of their involvement in a project and it is necessary to detect which addresses belong to the same person. The `find.aliases` function in the `tm.plugin.mail` package attempts to detect which email addresses refer to the same person.

### 15.6.3 Multiple sources of signals

Sometimes a value appearing in a sample could have come from multiple sources, only one of which is of interest. An example of this is the question: when did hexadecimal literals first appear as such in print?

One way of answering this question is to analyze the word n-grams (and associated year of book publication) Google have made available from their English book scanning project.<sup>518</sup>

The regular expression `^[\x{00}[\x{X}][\x{0-9a-fA-F}01]]` (ohh, Ohh and ell were treated as the corresponding digits) returned 89 thousand matches.

OCR mistakes have resulted in some words being treated as hexadecimal literals, e.g., Oxford scanned as 0xf0fd. The character sequence o xo is surprisingly common and looking at some of the contexts in which this occurs suggests that the usage is mainly related to chemical formula (some are also likely to be references to a product of that name).

Assuming that hexadecimal notation did not start appearing in books before electronic computers were invented, books prior to say 1945, the end of World War II, can be ignored; let's also ignore o xo.

It seems to me that if any hexadecimal literal appears in a book at least one more is likely to occur; applying this final filter reduced the number of matches to 7,292; with 319 unique character sequences.

Comparing the use of hexadecimal literals in C source with those extracted from Google books words we get:

### 15.6.4 Duplicate data

Duplicate data can occur for various reasons, including:

- collation of data from various sources, before it reaches the person performing the analysis, duplication of a particular list of values is always a possibility, e.g., all the values in one row/column match the values in another row/column or differ by a constant factor. This duplication may result in 100% correlation, e.g., one row contains temperature in Celsius while another uses Fahrenheit.
- repeated output from the measuring process. For instance, logging of computer faults where a single root cause can produce duplicate messages at sporadic times after the fault is experienced<sup>1334</sup> and spatially or functionally adjacent units to generate messages<sup>842</sup> (see [data-check/Blue-Gene.log](#)).

The `duplicated` function returns information about exact duplicate values appearing in rows.

When the data is numeric, *close* duplicates can be highlighted using pairwise correlation, e.g., Figure 8.7.

Some R functions handle duplicate row/columns gracefully (e.g., the `glm` function) while others give unpredictable results (e.g., the `solve` function which inverts a matrix), the behavior depends on the algorithm used and what if any consistency checks were added to the implementation by the person who wrote the code.

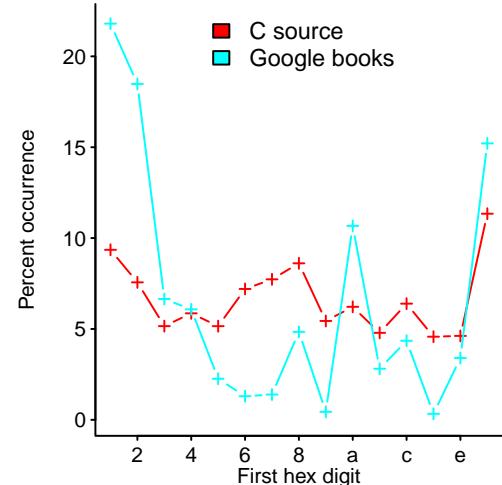


Figure 15.3: Percentage occurrence of the first digit of hexadecimal numbers in C source and estimated from Google book data. Data from Jones<sup>699</sup> and Michel et al.<sup>942</sup> code

### 15.6.5 Default values

Sometimes a measurement process returns what is considered to be a reasonable value if it cannot return the actual value. For instance, IP geolocation services are always able to associate a country with an IP address, but when they are unable to further refine the location within a country they return a location near the center of the country; for the USA this is close to the town of Potwin in Kansas (population 449) which appears to experience orders of magnitude more Internet related events for its population size than other towns in the US.<sup>678</sup>

### 15.6.6 Resolution limit of measurements

Some kinds of measurement are inherently inexact, e.g., time. When working close to the resolution limit of the measuring process, care needs to be taken to ensure that false signals are not generated by an interaction between the measurement resolution and the analysis procedure.

A study by Feitelson<sup>417</sup> measured the runtime of processes executed on a Unix based system to an accuracy of two decimal digits. Subsequent analysis of the number of processes whose execution fell within a given time interval showed a surprising result, there were many time intervals that did not contain any processes (upper plot in Figure 15.4). Subsequent analysis found that the timer resolution was 1/64 second and the gaps were an artefact of the number of digits recorded, recording more digits (lower plot in Figure 15.4) resulted in fewer intervals containing no measurement points.

## 15.7 Detecting fabricated data

A data set is not always derived from accurate measurements, the accuracy failures may be accidental or intentional, or might not involve any actual measurements (i.e., it has been fabricated).

Like all data analysis detection of fabricated data is based on finding known patterns in the data and as always the interpretation of why the data contains these patterns is the responsibility of the audience of the results; it is always worth repeating that domain knowledge is key.

One pattern of behavior that has been observed to occur with some regularity is that the first digit of sample values follows Benford's law to a reasonable degree of approximation (while a figure of 30% of all datasets has been quoted, the actual figure is likely to be much smaller<sup>1218</sup>). Benford's law has been used to detect accounting and election<sup>1180</sup> fraud, identification of fake survey interviews<sup>1214</sup> and scientific research.<sup>352</sup>

While references to Benford's law usually refer to the first digit of numeric values, there is a form that applies to the second and perhaps other significant digits.<sup>1010</sup> There has also been work<sup>113</sup> suggesting that the digit at the opposite end of numeric literals, the least significant digit, sometimes has a uniform distribution.

Benford's law says the probability of the first digit having value  $d$  is given by:

$$P(d) = \log_{10}(1 + \frac{1}{d})$$

Figure 7.33 investigates counts of the first digit of numeric literals in C source code.

If a set of independent and identically distributed random variables are sorted into order the distribution of digits of the differences between adjacent sorted values is close to Benford's law.<sup>953</sup> A test based on this fact can detect rounded data, data generated by linear regression and data generated by using the inverse function of a known distribution.<sup>1010</sup>

The `BenfordTests` package contains a variety of function for evaluating the conformity of a dataset to Benford's law.

When generating fabricated data, it is sometimes necessary to produce a random sequence of items. People hold incorrect beliefs<sup>151</sup> about the properties of random sequences and when asked to generate them produce sequences that contain predictable patterns (i.e., they are not random).

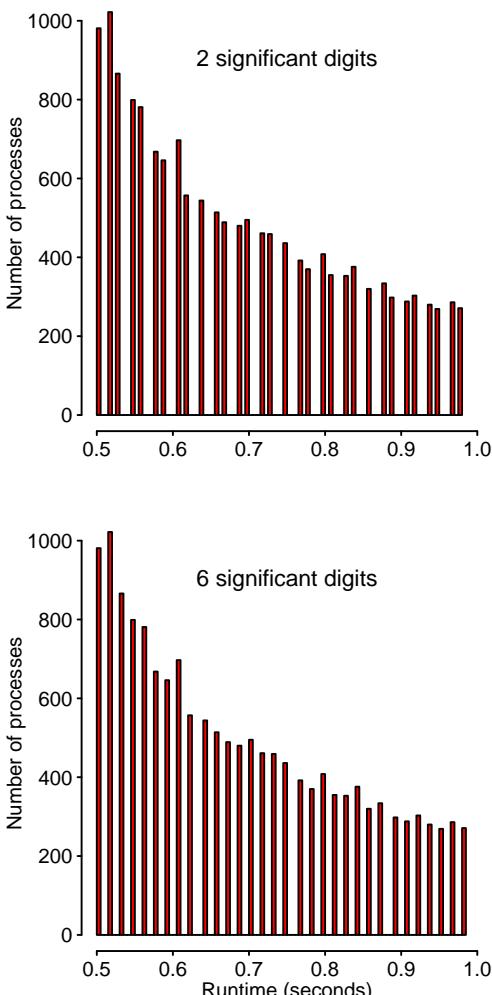


Figure 15.4: Number of processes executing for a given amount of time, with measurements expressed using two and six significant digits. Data from Feitelson.<sup>417</sup> [code](#)

One study<sup>1215</sup> was able to build a model that predicted repeated patterns in an individual's *randomly* selected sequence with around 25% success rate, but the model built for one person's behavior was used to make predictions about another persons the success rate dropped to around 18%.

Detecting divergence from or agreement with these patterns of behavior is dependent on the authors of the data set being lazy (i.e., being unwilling to spend the time making sure that the data they generate has the expected properties; the creators of the fictitious accounts publicly published by Madoff's companies, before his fraud was uncovered, made the effort to ensure that they followed Benford's law<sup>1222</sup>) or unfamiliar with the expected patterns of behavior.

#NoEstimates#NoEstimates, 107

# Bibliography

1. J. T. Abbott, J. L. Austerweil, and T. L. Griffiths. Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, 122(3):558–569, July 2015. [24](#)
2. T. Abdel-Hamid and S. E. Madnick. *Software Project Dynamics: An Integrated Approach*. Prentice-Hall, Inc, 1991. [107](#)
3. D. Aboody and B. Lev. The value relevance of intangibles: The case of software capitalization. *Journal of Accounting Research*, 36:161–191, 1998. [60](#)
4. Ada conformity assessment test suite (ACATS). website, Jan. 2018. <http://www.ada-auth.org/acats.html>. [146](#)
5. E. N. Adams. Optimizing preventive service of software products. *IBM Journal of Research and Development*, 28(1):2–14, Jan. 1984. [133](#)
6. J. Adams. *Risk and Freedom: The record of road safety regulation*. Transport Publishing Projects, 1985. [44](#)
7. B. Adelson. Problem solving and the development of abstract categories in programming languages. *Memory & Cognition*, 9(4):422–433, July 1981. [34](#), [35](#)
8. J. Agar. *The Government Machine A Revolutionary History of the Computer*. The MIT Press, 2003. [67](#)
9. R. Agarwal and M. Gort. The evolution of markets and entry, exit and survival of firms. *The Review of Economics and Statistics*, 78(3):489–498, Aug. 1996. [79](#)
10. P. J. Ågerfalk. Insufficient theoretical contribution: a conclusive rationale for rejection? *European Journal of Information Systems*, 23(6):593–599, Nov. 2014. [5](#)
11. A. Aghayev and P. Desnoyers. Skylight-A window on shingled disk operation. In *13th USENIX Conference on File and Storage Technologies*, FAST’15, pages 135–149, Feb. 2015. [340](#)
12. N. Agrawal. *Representative, reproducible, and practical benchmarking of file and storage systems*. PhD thesis, University of Wisconsin-Madison, 2009. [321](#)
13. N. Agrawal, A. C. Arpaci-Dusseau, and R. H. Arpaci-Dusseau. Generating realistic impressions for file-system benchmarking. *ACM Transactions on Storage*, 5(4):125–138, Dec. 2009. [190](#)
14. N. Agrawal, W. J. Bolosky, J. R. Douceur, and J. R. Lorch. A five-year study of file-system metadata. *ACM Transactions on Storage*, 3(3):31–45, Oct. 2007. [211](#)
15. J. J. Ahonen and P. Savolainen. Software engineering projects may fail before they are started: Post-mortem analysis of five cancelled projects. *Journal of Systems and Software*, 83(11):2175–2187, Nov. 2010. [102](#)
16. G. A. Akerlof. The market for "Lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, Aug. 1970. [57](#)
17. K. Akita, S. Itagaki, Y. Masawa, M. Nonaka, T. Hatani, K. Hattori, S. Morisaki, Y. Yanagida, T. Takaya, T. Furuyama, and O. Takashi. *Software Development Data White paper 2012-2013*. SEC BOOKS, 2012. [101](#)
18. H. A. A. Al-Mutawa. On the classification of cyclic dependencies in Java programs. Thesis (m.s.), Massey University, New Zealand, 2013. [93](#)
19. H. Alemzadeh, R. K. Iyer, Z. Kalbarczyk, and J. Raman. Analysis of safety-critical computer failures in medical devices. *IEEE Security & Privacy*, 11(4):14–26, July 2013. [250](#), [251](#), [255](#)
20. N. Ali, Z. Sharafi, Y.-G. Guéhéneuc, and G. Antoniol. An empirical study on requirements traceability using eye-tracking. In *28th IEEE International Conference on Software Maintenance*, ICSM 2012, pages 191–200, Sept. 2012. [35](#)
21. R. C. Allen. The British industrial revolution in global perspective: How commerce created the industrial revolution and modern economic growth. *unpublished???, ???(???)::???, Apr. 2007.* [67](#)
22. L. Allodi. Economic factors of vulnerability trade and exploitation. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS’17, pages 1483–1499, Oct.-Nov. 2017. [129](#)
23. L. Allodi and F. Massacci. A preliminary analysis of vulnerability scores for attacks in wild: The EKITS and SYM datasets. In *Proceedings of the 2012 ACM Workshop on Building analysis datasets and gathering experience returns for security*, DABGERS’12, pages 17–24, Oct. 2012. [129](#)
24. M. G. Almiron, E. S. Almeida, and M. N. Miranda. The reliability of statistical functions in four software packages freely used in numerical computation. *Brazilian Journal of Probability and Statistics*, 23(2):107–119, 2009. [11](#)
25. M. G. Almiron, B. Lopes, A. L. C. Oliveira, A. C. Medeiros, and A. C. Frery. On the numerical accuracy of spreadsheets. *Journal of Statistics*, 34(4):1–29, Apr. 2010. [11](#)
26. L. E. Altenereder. The learning curve in solving a jig-saw puzzle: A teaching device. *Journal of Educational Psychology*, 26(3):231–232, Mar. 1935. [24](#), [25](#)
27. E. M. Altmann. *Episodic Memory for External Information*. PhD thesis, Carnegie Mellon University, Aug. 1996. [20](#)
28. E. M. Altmann. Functional decay of memory for tasks. *Psychological Research*, 66(4):287–297, 2002. [44](#)
29. E. M. Altmann, J. G. Trafton, and D. Z. Hambrick. Effects of interruption length on procedural errors. *Journal of Experimental Psychology: Applied*, 23(2):216–229, June 2017. [23](#)
30. H. Aman, S. Amasaki, T. Yokogawa, and M. Kawahara. A survival analysis of source files modified by new developers. In *International Conference on Product-Focused Software Process Improvement*, PROFES 2017, pages 80–88, Nov.-Dec. 2017. [300](#)
31. Amazon, Inc. Amazon ec2 service level agreement. <https://aws.amazon.com/ec2/sla/>, June 2013. [345](#)
32. S. Ambler. IT project success survey results. <http://www.ambysoft.com/surveys/>, 2017. [102](#)
33. J. M. Amiri and V. V. K. Padmanabhu. A comprehensive evaluation of conversion approaches for different function points. Thesis (m.s.), Blekinge Institute of Technology, Sweden, Sept. 2011. [251](#)
34. L. An, O. Mlouki, F. Khomh, and G. Antoniol. Stack overflow: A code laundering platform? In *eprint arXiv:cs/1703.03897*, Mar. 2017. [96](#)
35. L. V. B. An T. Oskarsson, G. H. McClelland, and R. Hastie. What’s next? Judging sequences of binary events. *Psychological Bulletin*, 135(2):262–285, 2009. [17](#), [137](#)
36. They’re (almost) all dirty: The state of cheating in Android benchmarks. website, Oct. 2013. <http://www.anandtech.com/show/7384/state-of-cheating-in-android-benchmarks>. [336](#)
37. B. C. D. Anda, D. I. K. Sjøberg, and A. Mockus. Variability and reproducibility in software engineering: A study of four companies that developed the same system. *IEEE Transactions on Software Engineering*, 35(3):407–429, May 2009. [102](#), [109](#)
38. J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, Apr. 2007. [17](#)
39. J. R. Anderson and R. Milson. Human memory: An adaptive perspective. *Psychological Review*, 96(4):703–719, 1989. [23](#)
40. M. L. Anderson. Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4):245–313, Apr. 2010. [15](#)
41. GNU/Linux distribution timeline. website, Oct. 2012. <http://futurist.se/gldt>. [87](#)
42. The short long. Speech, May 2011. 29th Société Universitaire Européene de Recherches Financières Colloquium. [79](#)
43. D. Andriesse, X. Chen, V. van der Veen, A. Slowinska, and H. Bos. An in-depth analysis of disassembly on full-scale x86/x64 binaries. In *Proceedings of the 25th USENIX Security Symposium*, pages 583–600, Aug. 2016. [146](#)

44. J. Annett. Subjective rating scales: science or art? *Ergonomics*, 45(12):966–987, 2002. 346
45. A. Ansar. 'AppStore secrets' (What we've learned from 30,000,000 downloads). Presentation, pinch media, Aug. 2017. 84
46. F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, Feb. 1973. 251, 252
47. K. Aoki and M. W. Feldman. Evolution of learning strategies in temporally and spatially variable environments: A review of theory. *Theoretical Population Biology*, 91:3–19, Feb. 2014. 71
48. J. Aranda. Anchoring and adjustment in software estimation. Thesis (m.s.), Graduate Department of Computer Science, University of Toronto, 2005. 105
49. H. R. Arkes, R. M. Dawes, and C. Christensen. Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, 37:93–110, 1986. 48
50. P. Armer. SHARE – A eulogy to cooperative effort. Technical Report P-969, The RAND Corporation, Oct. 1956. 70
51. J. S. Armstrong. The seer-sucker theory: The value of experts in forecasting. *Technology Review*, pages 16–24, June-July 1980. 75
52. V. Arnaoudova, L. M. Eshkevari, M. Di Penta, R. Oliveto, G. Antoniol, and Y.-G. Guéhéneuc. REPENT: Analyzing the nature of identifier renamings. *IEEE Transactions on Software Engineering*, 40(5):502–532, May 2014. 163
53. J. Arndt. *Matters Computational: Ideas, Algorithms, Source Code*. Springer, 2010. 169
54. T. B. Arnold and J. W. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39, Dec. 2011. 204
55. A. Arora, R. Krishnan, R. Telang, and Y. Yang. An empirical analysis of software vendors' patch release behavior: Impact of vulnerability disclosure. *Information Systems Research*, 21(1):115–132, Mar. 2010. 298, 300, 364
56. W. B. Arthur. *Increasing Returns and Path Dependency in the Economy*. The University of Michigan Press, 1994. 97
57. S. E. Asch. Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, 70(9):1–70, 1956. 47
58. T. A. Åstebro, S. A. Jeffrey, and G. K. Adomdza. Inventor perseverance after being told to quit: The role of cognitive biases. *Journal of Behavioral Decision Making*, 20(3):253–272, Apr. 2007. 48
59. S. Atkinson and G. Benefield. Software development: Why the traditional contract model is not fit for purpose. In *46th Hawaii International Conference on System Sciences*, HICSS, pages 4842–4851, Jan. 2013. 110
60. V. Atlidakis, J. Andrus, R. Geambasu, D. Mitropoulos, and J. Nieh. POSIX abstractions in modern operating systems: The old, the new, and the missing. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys'16, page 19, Apr. 2016. 91
61. Audit Scotland. i6: a review. Report, Audit Scotland, Mar. 2017. 102
62. Auerbach. Auerbach guide to time sharing. Computer technology report, Auerbach Publishers Inc., Jan. 1973. 77
63. N. R. Augustine. *Augustine's Laws*. American Institute of Aeronautics and Astronautics, Inc, sixth edition, 1997. 130
64. R. Auler and E. Borin. A LLVM just-in-time compilation cost analysis. Technical Report IC-13-13, Instituto de Computação Universidade Estadual de Campinas, May 2013. 151, 152
65. J. Autran, D. Munteanu, P. Roche, and G. Gasior. Real-time soft-error rate measurements: A review. *Microelectronics Reliability*, 54(8):1455–1476, Aug. 2014. 142
66. J.-L. Autran, S. Semikh, D. Munteanu, S. Serre, G. Gasior, and P. Roche. Soft-error rate of advanced SRAM memories: Modeling and monte carlo simulation. In M. Andriyuchuk, editor, *Numerical Simulation – From Theory to Industry*, chapter 15, pages 309–336. InTech, Sept. 2012. 142
67. E. Avidan. The significance of method parameters and local variables as beacons for comprehension: An empirical study. Thesis (m.s.), The Hebrew University of Jerusalem, Nov. 2016. 163
68. P. Azoulay, C. Fons-Rosen, and J. S. G. Zivin. Does science advance one funeral at a time? Working Paper No. 21788, National Bureau of Economic Research, Dec. 2015. 7
69. R. H. Baayen, P. Milin, and M. Ramscar. Frequency in lexical processing. *Aphasiology*, 30(11):1174–1220, Mar. 2016. 161
70. V. Babka. *Improving Accuracy of Software Performance Models on Multicore Platforms with Shared Caches*. PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague, Oct. 2012. 342
71. V. Babka and P. Táma. Investigating cache parameters of x86 family processors. In *Proceedings of the 2009 SPEC Benchmark Workshop on Computer Performance Evaluation and Benchmarking*, pages 77–96, Jan. 2009. 343
72. D. Baccarini, G. Salm, and P. E. D. Love. Management of risks in information technology projects. *Industrial Management & Data Systems*, 104(4):286–295, 2004. 111
73. A. Bacchelli and C. Bird. Expectations, outcomes, and challenges of modern code review. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE'13, pages 712–721, May 2013. 145
74. A. Bachmann, C. Bird, F. Rahman, P. Devanbu, and A. Bernstein. The missing links: Bugs and bug-fix commits. In *Proceedings of the 18th ACM SIGSOFT international symposium on Foundations of software engineering*, FSE 2010, pages 97–106, Nov. 2010. 127
75. J. Backus. The history of FORTRAN I, II, and III. *SIGPLAN Notices*, 13(8):165–180, 1978. 88
76. J. Backus. Programming in America in the 1950s—some personal impressions. In N. Metropolis, J. Howlett, and G.-C. Rota, editors, *A History of Computing in the Twentieth Century*, pages 125–135. Academic Press, Inc, Feb. 1981. 70
77. J. W. Backus, R. J. Beeber, S. Best, R. Goldberg, H. Herrick, R. A. Hughes, L. B. Mitchell, R. A. Nelson, R. Nutt, D. Sayre, P. B. Sheridan, H. Stern, and I. Ziller. *The FORTRAN Automatic Coding System for the IBM 704 EDPM: Programmer's Reference Manual*. International Business Machines Corporation, 590 Madison Avenue, New York 22, N.Y., Oct. 1956. 89
78. A. Bacon, S. Handley, and S. Newstead. Individual differences in strategies for syllogistic reasoning. *Thinking & Reasoning*, 9(2):133–168, 2003. 38
79. A. Baddeley. Working memory. In A. Baddeley, M. W. Eysenck, and M. Anderson, editors, *Memory*, chapter 3, pages 41–69. Psychology Press, Feb. 2009. 21
80. A. Baddeley. Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63:1–29, Sept. 2012. 21
81. A. D. Baddeley, N. Thomson, and M. Buchanan. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6):575–589, Dec. 1975. 21, 322
82. M. Bagherzadeh, N. Kahani, C.-P. Bezemer, A. E. Hassan, J. Dingel, and J. R. Cordy. Analyzing a decade of Linux system calls. *Empirical Software Engineering*, 23(3):1519–1551, June 2018. 92
83. J. N. Bailenson, M. S. Shum, S. Atran, D. L. Medin, and J. D. Coley. A bird's eye view: biological categorization and reasoning within and across cultures. *Cognition*, 84:1–53, 2002. 30
84. D. H. Bailey. Misleading performance reporting in the supercomputer field. Technical Report RNR-92-005, Numerical Aerodynamic Simulation Division, NASA Ames Research Center, Dec. 1992. 336
85. S. Baily, R. Gilbertson, and E. Straub. Modular multimode radar (CMMR) software acquisition study. Technical Report 2302-01-1-2291, ARINC Research Corporation, Mar. 1981. 79
86. E. Bainomugisha, A. L. Carreton, T. van Cutsem, S. Mostinckx, and W. de Meuter. A survey on reactive programming. *ACM Computing Surveys*, 45(4):52, Aug. 2013. 150
87. P. Bajari, S. Tadelis, and S. Houghton. Bidding for incomplete contracts: An empirical analysis of adaptation costs. *American Economic Review*, 101(4):1288–1319, Oct. 2011. 109
88. F. T. Baker. Chief programmer team management of production programming. *IBM Systems Journal*, 11(1):56–73, 1972. 121
89. M. Bakkaloglu, J. J. Wylie, C. Wang, and G. R. Ganger. On correlated failures in survivable storage systems. Technical Report CMU-CS-02-129, Carnegie Mellon University, May 2002. 334
90. B. Balaji, J. McCullough, R. K. Gupta, and Y. Agarwal. Accurate characterization of the variability in power consumption in modern mobile processors. In *Proceedings of the 2012 USENIX conference on Power-Aware Computing and Systems (HotPower'12)*, Oct. 2012. 277, 278

91. N. Banerjee, A. Rahmati, M. D. Corner, S. Rollins, and L. Zhong. Users and batteries: Interactions and adaptive energy management in mobile systems. In *International Conference on Ubiquitous Computing*, UbiComp 2007, pages 217–237, Sept. 2007. [78](#)
92. P. Banyard and N. Hunt. Something missing? *The Psychologist*, 13(2):68–71, 2000. [17](#)
93. J. H. Barkow, L. Cosmides, and J. Tooby. *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford University Press, 1992. [16](#)
94. W. P. Barnett. *The Red Queen among Organizations: How competitiveness evolves*. Princeton University Press, 2008. [58](#)
95. A. Baronchelli, V. Loreto, and A. Puglisi. Individual biases, cultural evolution, and the statistical nature of language universals: The case of colour naming systems. *PLoS ONE*, 10(5):e0125019, May 2015. [30](#)
96. L. Barrett, R. Dunbar, and J. Lycett. *Human Evolutionary Psychology*. Palgrave Macmillan, 2002. [16](#)
97. L. A. Barroso and U. Hölzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines. Report, Morgan & Claypool, 2009. [78](#)
98. V. R. Basili, N. M. P.-Y. amd Connie Loggia Ramsey, C. Shih, and E. E. Katz. A quantitative analysis of software developed in Ada. Technical Report TR-1403, Department of Computer Science, University of Maryland, May 1984. [137](#)
99. V. R. Basili and J. Beane. Can the Parr curve help with manpower distribution and resource estimation problems? *The Journal of Systems and Software*, 2(1):59–69, Feb. 1981. [107](#)
100. V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, and M. V. Zelkowitz. The empirical investigation of perspective-based reading. In *Proceedings of the Twentieth Annual Software Engineering Workshop*, pages 21–69, Dec. 1995. [6](#), [319](#)
101. V. R. Basili and A. J. Turner. Iterative enhancement: A practical technique for software development. *IEEE Transactions on Software Engineering*, SE-1(4):390–396, Dec. 1975. [114](#)
102. F. M. Bass. A new product growth model for consumer durables. *Management Science*, 15(5):215–227, Jan. 1969. [62](#)
103. H. A. Bastiaanse. *Very, Many, Small, Penguins: Vaguely Related Topics*. PhD thesis, Institute for Logic, Language and Computation, Universiteit van Amsterdam, Mar. 2014. [137](#)
104. B. Baudry, S. Allier, and M. Monperrus. Tailored source code transformations to synthesize computationally diverse program variants. In *eprint arXiv:cs.SE/1401.7635v1*, Jan. 2014. [139](#), [165](#)
105. F. L. Bauer and H. Wössner. The "Plankalkül" of Konrad Zuse: a forerunner of today's programming languages. *Communications of the ACM*, 15(7):678–685, July 1972. [89](#)
106. A. Baumann. Hardware is the new software. In *Proceedings of the 16th Workshop on Hot Topics in Operating Systems*, HotOS'17, pages 132–137, May 2016. [88](#)
107. R. F. Baumeister. *Is There Anything Good About Men?* Oxford University Press, 2010. [17](#)
108. R. T. Baust. *Computer Characteristics Quarterly: Volume 7, Number 4-Volume 8, Number 1*. adams associates, 1968. [78](#)
109. G. Bavota, G. Canfora, M. Di Penta, R. Oliveto, and S. Panichella. The evolution of project inter-dependencies in a software ecosystem: the case of Apache. In *Proceedings of the 2013 IEEE International Conference on Software Maintenance*, ICSM'13, pages 280–289, Sept. 2013. [121](#)
110. G. Bavota, G. Canfora, M. Di Penta, R. Oliveto, and S. Panichella. How the Apache community upgrades dependencies: An evolutionary study? *Empirical Software Engineering*, 20(5):1275–1317, Oct. 2015. [121](#)
111. O. Baysal, O. Kononenko, R. Holmes, and M. W. Godfrey. The influence of non-technical factors on code review. In *20th Working Conference on Reverse Engineering*, WCRE 2013, pages 122–131, Oct. 2013. [86](#)
112. B. L. Bayus, S. Jain, and A. G. Rao. Truth or consequences: An analysis of vaporware and new product announcements. *Journal of Marketing Research*, 38(1):3–13, Feb. 2001. [65](#)
113. B. Beber and A. Scacco. What the numbers say: A digit-based test for election fraud. *Political Analysis*, 20(2):211–234, Apr. 2012. [372](#)
114. G. S. Becker. Investment in human capital: A theoretical analysis. *Journal of Political Economy*, 70(5):9–49, Oct. 1962. [82](#)
115. R. A. Becker and W. S. Cleveland. *Trellis Graphics User's Manual*. AT&T Bell Laboratories, Murray Hill, Dec. 1995. [189](#)
116. J. Beckhusen. Occupations in information technology. American Community Survey Report ACS-35, U.S. Census Bureau, Aug. 2016. [81](#)
117. A. Begel and T. Zimmermann. Analyze this! 145 questions for data scientists in software engineering. Technical Report MSR-TR-2013-111, Microsoft Research, Oct. 2013. [5](#), [346](#)
118. M. Bekoff, C. Allen, and G. M. Burghardt. *The Cognitive Animal: Empirical and Theoretical Perspectives on Animal Cognition*. MIT Press, 2002. [16](#)
119. C. G. Bell. Fundamentals of time shared computers. *Computer Design*, 7(2):44–59, Feb. 1968. [1](#)
120. G. Bell. Supercomputers: The amazing race (A history of supercomputing, 1960–2020). Technical Report MSR-TR-2015-2, Microsoft Research, Microsoft Corporation, Nov. 2014. [78](#)
121. V. A. Bell and P. N. Johnson-Laird. A model theory of modal reasoning. *Cognitive Science*, 22(1):25–51, 1998. [36](#), [37](#)
122. M. Beller, A. Zaidman, A. Karpov, and R. A. Zwaan. The last line effect explained. *Empirical Software Engineering*, 22(3):1508–1536, June 2017. [96](#), [139](#)
123. R. W. Bemer. A view of the history of COBOL. *Honeywell Computer Journal*, 5(3):130–135, Nov. 1959. [88](#)
124. O. A. Ben-Yehuda, M. Ben-Yehuda, A. Schuster, and D. Tsafir. Deconstructing Amazon EC2 spot instance pricing. *ACM Transactions on Economics and Computation*, 1(3), Sept. 2013. [54](#)
125. G. Beniamini, S. Gingichashvili, A. K. Orbach, and D. G. Feitelson. Meaningful identifier names: The case of single-letter variables. In *25th IEEE International Conference on Program Comprehension*, ICPC 2017, pages 45–54, May 2017. [73](#)
126. Y. Benkler. Coase's penguin, or, Linux and the nature of the firm. *The Yale Law Journal*, 112(3), Dec. 2002. [53](#), [64](#)
127. F. C. Y. Benureau and N. P. Rougier. Re-run, repeat, reproduce, reuse, replicate: Transforming code into scientific contributions. In *eprint arXiv:cs.GL/1708.08205*, Aug. 2017. [91](#)
128. T. Berger, S. She, K. Czarnecki, and A. Wąsowski. Feature-to-code mapping in two large product lines. In J. Bosch and J. Lee, editors, *Software Product Lines: Going Beyond*, volume 6287 of *Lecture Notes in Computer Science*, pages 498–499. Springer Berlin Heidelberg, 2010. [205](#)
129. T. Berger, S. She, R. Lotufo, A. Wąsowski, and K. Czarnecki. Variability modeling in the systems software domain. Technical Report GSMLAB-TR 2012-07-06, Generative Software Development Laboratory, University of Waterloo, July 2012. [168](#), [255](#)
130. E. Berghout, M. Nijland, and K. Grant. Seven ways to get your favoured IT project accepted – politics in IT evaluation. *The Electronic Journal of Information Systems Evaluation*, 8(1):31–40, 2005. [102](#)
131. B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. Berkeley: University of California Press, 1969. [30](#)
132. D. Bermbach and E. Wittern. Benchmarking web API quality. In *2016 International Conference on Web Engineering*, ICWE'16, pages 188–206, June 2016. [143](#)
133. A. Bernardo and I. Welch. On the evolution of overconfidence and entrepreneurs. *Journal of Economics & Management Strategy*, 10(3):301–330, 2001. [72](#)
134. K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson, and N. J. Rohrer. High-performance CMOS variability in the 65-nm regime and beyond. *IBM Journal of Research and Development*, 50(4/5):433–449, July 2006. [336](#)
135. D. M. Berry, E. Kamsties, and M. M. Krieger. From contract details to software specification: Linguistic sources of ambiguity. Nov. 2003. [138](#)
136. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In C. Beeri and P. Buneman, editors, *Database Theory-ICDT'99: 7th International Conference*, pages 217–235. Springer-Verlag, Jan. 1999. [309](#)
137. D. Biber, S. Johansson, G. Leech, S. Conrad, and E. Finegan. *Longman Grammar of Spoken and Written English*. Pearson Education, 1999. [37](#), [137](#), [164](#)

138. B. Biegel, F. Beck, W. Hornig, and S. Diehl. The order of things: How developers sort fields and methods. In *28th IEEE International Conference on Software Maintenance*, ICSM, pages 88–97, Sept. 2012. [171](#), [172](#), [314](#), [315](#)
139. S. Bikchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, Oct. 1992. [47](#)
140. P. Bilton, P. Dodimead, E. Livingstone, I. Rayner, G. Turner, M. Wynniatt, and S. Howes. Managing the risks of legacy ICT to public service delivery. HC 539 SESSION 2013-14, National Audit Office, UK, Sept. 2013. [77](#), [85](#)
141. W. L. Bircher. *Predictive Power Management for Multi-Core Processors*. PhD thesis, The University of Texas at Austin, Dec. 2010. [338](#), [341](#)
142. C. Bird, A. Bachmann, E. Aune, J. Duffy, A. Bernstein, V. Filkov, and P. Devanbu. Fair and balanced? Bias in bug-fix datasets. In *Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, FSE 2009, pages 121–130, Aug. 2009. [127](#)
143. S. Bird. Software knows best: A case for hardware transparency and measurability. Thesis (m.s.), Department of Electrical Engineering and Computer Science, University of California at Berkeley, May 2010. [321](#)
144. P. G. Bishop and R. E. Bloomfield. Worst case reliability prediction based on a prior estimate of residual defects. In *Proceedings 13th International Symposium on Software Reliability Engineering*, ISSRE'02, pages 295–303, Nov. 2002. [134](#)
145. T. F. Bissyandé, F. Thung, D. Lo, L. Jiang, and L. Réveillère. Popularity, interoperability, and impact of programming languages in 100,000 open source projects. In *37th Annual International Computer Software & Applications Conference*, COMPSAC 2013, pages 303–312, July 2013. [91](#), [184](#)
146. Bitsavers’ pdf document archive. website, June 2017. <http://bitsavers.trailing-edge.com/pdf>. [91](#)
147. N. M. Blachman. A survey of automatic digital computers. Survey 111293, Office of Naval Research, Washington, D.C., 1953. [336](#)
148. S. M. Blackburn, R. Garner, C. Hoffmann, A. M. Khan, K. S. McKinley, R. Bentzur, A. Diwan, D. Feinberg, D. Frampton, S. Z. Guyer, M. Hirzel, A. Hosking, M. Jump, H. Lee, J. E. B. Moss, A. Phansalkar, D. Stefanović, T. VanDrunen, D. von Dincklage, and B. Wiedermann. The DaCapo benchmarks: Java benchmarking development and analysis (extended version). Technical Report TR-CS-06-01, Department of Computer Science, Australian National University, Aug. 2006. [325](#)
149. A.-R. Blais and E. U. Weber. A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1):33–47, Apr. 2006. [44](#)
150. M. S. Blaubergs and M. D. S. Braine. Short-term memory limitations on decoding self-embedded sentences. *Journal of Experimental Psychology*, 102(4):745–748, 1974. [22](#)
151. D. S. Blinder and D. M. Oppenheimer. Beliefs about what types of mechanisms produce random sequences. *Journal of Behavioral Decision Making*, 21(4):414–427, Oct. 2008. [372](#)
152. B. W. Boehm. *Software Engineering Economics*. Prentice-Hall, Inc, 1981. [251](#)
153. G. D. Boetticher. Improving credibility of machine learner models in software engineering. In D. Zhang and J. J. P. Tsai, editors, *Advances in Machine Learning Applications in Software Engineering*, chapter 3, pages 52–73. Idea Group Publishing, Oct. 2006. [364](#)
154. A. Bohn, I. Feinerer, K. Hornik, and P. Mair. Content-based social network analysis of mailing lists. *The R Journal*, 3(1):11–18, June 2011. [316](#), [370](#)
155. J. G. Bolten, R. S. Leonard, M. V. Arena, O. Younossi, and J. M. Sollinger. Sources of weapon system cost growth: Analysis of 35 major defense acquisition programs. Monograph series, RAND Corporation, 2008. [105](#)
156. R. Bornat, S. Dehnadi, and S. ??? Mental models, consistency and programming aptitude. In *Tenth Australasian Computing Education Conference*, ACE'08, pages 53–61, Jan. 2008. [150](#)
157. L. Boroditsky. Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, 75:1–28, 2000. [73](#)
158. A. Börsch-Supan and M. Weiss. Productivity and age: Evidence from work teams at the assembly line. Technical Report 148-2007, Manheim Research Institute for the Economics of Aging, 2007. [49](#)
159. L. Bossavit. *The Leprechauns of Software Engineering: How folklore turns into fact and what to do about it*. Leanpub, 2016. [59](#)
160. N. Bostrom and A. Sandberg. The wisdom of nature: An evolutionary heuristic for human enhancement. In J. Savulescu and N. Bostrom, editors, *Human Enhancement*, chapter 18, pages 375–416. Oxford University Press, Jan. 2011. [16](#)
161. A. Botchkarev. Estimating the accuracy of the return on investment (ROI) performance evaluations. *Interdisciplinary Journal of Information, Knowledge, and Management*, 10:217–233, 2015. [54](#)
162. J. Bourn. New IT systems for Magistrates’ courts: the Libra project. Report by the Comptroller and Auditor General HC 327 Session 2002-2003, National Audit Office, Jan. 2003. [110](#)
163. G. H. Bower, J. B. Black, and T. J. Turner. Scripts in memory for text. *Cognitive Psychology*, 11(2):177–220, Apr. 1979. [156](#)
164. J. S. Bowers and C. J. Davis. Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, 138(3):389–414, 2012. [220](#)
165. R. Boyd and P. J. Richerson. Why culture is common, but cultural evolution is rare. *Proceedings of the British Academy*, 88(???):77–93, Apr. 1988. [71](#)
166. R. Boyd and P. J. Richerson. Why does culture increase human adaptability. *Ethology and Sociobiology*, 16(2):125–143, Mar. 1995. [71](#)
167. M. G. Bradac, D. E. Perry, and L. G. Votta. Prototyping a process monitoring experiment. *IEEE Transactions on Software Engineering*, 20(10):774–784, 1994. [114](#), [115](#)
168. T. F. Brady, T. Konkle, G. A. Alvarez, and A. Oliva. Visual long-term memory has a massive storage capacity for object details. *PNAS*, 105(38):14325–14329, Sept. 2008. [50](#)
169. D. Braha and Y. Bar-Yam. The statistical mechanics of complex product development: Empirical and analytical results. *Management Science*, 53(7):1127–1145, July 2007. [150](#), [151](#)
170. D. W. Braithwaite and R. L. Goldstone. Flexibility in data interpretation: effects of representational format. *Frontiers in Psychology*, 4(980):1–16, Dec. 2013. [188](#)
171. M. C. Branco, Y. Xiong, K. Czarnecki, J. Küster, and H. Völzer. An empirical study on consistency management of business and IT process models. Technical Report GSMLAB-TR 2012-03-02, Generative Software Development Laboratory, University of Waterloo, Mar. 2012. [69](#)
172. S. Brand. *How buildings Learn: What happens after they’re built*. Viking, 1994. [112](#)
173. J. D. Bransford and J. J. Franks. The abstraction of linguistic ideas. *Cognitive Psychology*, 2(4):331–350, Oct. 1971. [155](#), [156](#)
174. J. D. Bransford and M. K. Johnson. Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11(6):717–726, Dec. 1972. [154](#), [155](#)
175. S. Brass and C. Goldberg. Semantic errors in SQL queries: A quite complete list. *Journal of Systems and Software*, 79(5):630–644, May 2006. [152](#)
176. R. A. Brealey, S. C. Myers, and F. Allen. *Principles of Corporate Finance*. McGraw-Hill Irwin, 10th edition, 2011. [56](#)
177. B. Brembs, K. Button, and M. Munafò. Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience*, 7(291), June 2013. [8](#)
178. S. Breu, R. Premraj, J. Sillito, and T. Zimmermann. Information needs in bug reports: Improving cooperation between developers and users. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW’10, pages 301–310, Feb. 2010. [183](#), [184](#)
179. C. A. Brewer. Color use guidelines for mapping and visualization. In A. M. MacEachren and D. R. F. Taylor, editors, *Visualization in Modern Cartography*, chapter 7, pages 123–147. Pergamon, Nov. 1994. [193](#)
180. E. Brewer, L. Ying, L. Greenfield, R. Cypher, and T. Ts'o. Disks for data centers. Technical report, Google, Inc, Feb. 2016. [340](#)
181. Brigham Young trace repository. No longer available via Brigham Young website, 201? Copy kindly supplied by Dror G. Feitelson. [131](#)

182. P. Brinch Hansen and R. House. The COBOL compiler for the Siemens 3003. *BIT*, 6(1):1–23, Mar. 1966. 89
183. S. Broadbent. Font requirements for next generation air traffic management systems. Technical Report HRS/HSP-006-REP-01, European Organisation for the Safety of Air Navigation, 2000. 35
184. G. W. Brock. *The U.S. Computer Industry: A Study of Market Power*. Ballinger Publishing Company, 1975. 68
185. L. D. Brock and H. A. Goodman. Reliability analysis of the F-8 digital fly-by-wire system. NASA Contractor Report 163110, Dryden Flight Research Center, Oct. 1981. 123
186. G. Bronevetsky and B. R. de Supinski. Soft error vulnerability of iterative linear algebra methods. In *Proceedings of the 22nd Annual International Conference on Supercomputing*, ICS '08, pages 155–164, June 2008. 142
187. J. Brooke. SUS: A 'quick' and 'dirty' usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, editors, *Usability Evaluation in Industry*, chapter 21, pages 189–194. Taylor and Francis, June 1996. 345
188. J. Brooke. SUS: A retrospective. *Journal of Usability Studies*, 8(2):29–40, Feb. 2013. 345
189. R. Brooks. A model of human cognitive behavior in writing code for computer programs, vol I. Report AFOSR-TR-75-1084, Carnegie-Mellon University, May 1975. 25
190. F. P. Brooks, Jr. *The Mythical Man-Month*. Addison-Wesley, anniversary edition, 1995. 6, 122
191. G. D. A. Brown, I. Neath, and N. Chater. A temporal ratio model of memory. *Psychological Review*, 114(3):539–576, 2007. 23
192. J. Brunner and P. C. Austin. Inflation of Type I error rate in multiple regression when independent variables are measured with error. *The Canadian Journal of Statistics*, 37(1):33–46, Mar. 2009. 244
193. M. Brysbaert, W. Fias, and M.-P. Noël. The Whorfian hypothesis and numerical cognition: is 'twenty-four' processed in the same way as 'four-and-twenty'? *Cognition*, 66(1):51–77, Apr. 1998. 166
194. I. Buchmann. *Batteries in a Portable World: A Handbook on rechargeable Batteries for Non-engineers*. Cadex Electronix Inc, third edition, 2011. 339
195. J. B. Buckheit and D. L. Donoho. WaveLab and reproducible research. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, chapter 5, pages 55–81. Springer-Verlag, 1995. 7
196. M. Budden, P. Hadavas, L. Hoffman, and C. Pretz. Generating valid  $4 \times 4$  correlation matrices. *Applied Mathematics E-Notes*, 7:53–59, 2007. 199
197. D. V. Budescu, H.-H. Por, S. B. Broomell, and M. Smithson. The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, 4:508–512, Apr. 2014. 127, 196
198. D. J. Buettner. *Designing an Optimal Software Intensive System Acquisition: A Game Theoretic Approach*. PhD thesis, University of Southern California, Sept. 2008. 107, 113, 121, 286, 292, 315
199. M. Bullynck. What is an operating system? A historical investigation (1954–1964). HAL Id: halshs-01541602, HAL archives-ouvertes.fr, Aug. 2017. 70
200. N. Bulnet and M. H. Halstead. Impurities found in algorithm implementations. Technical Report CSD-TR 111, Purdue University, Mar. 1974. 164
201. J. S. Bunderson and K. M. Sutcliffe. Management team learning orientation and business unit performance. *Journal of Applied Psychology*, 88(3):552–560, June 2003. 76
202. K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33(2):261–304, Nov. 2004. 246
203. Q. L. Burrell. A note on ageing in a library circulation model. *Journal of Documentation*, 41(2):100–115, 1985. 23
204. R. P. L. Buse and W. R. Weimer. Learning a metric for code readability. *IEEE Transactions on Software Engineering*, 36(4):546–558, July 2010. 163
205. J. Businge. *Co-evolution of the Eclipse Framework and its Third-party Plug-ins*. PhD thesis, Eindhoven University of Technology, Sept. 2013. 296, 297
206. J. Businge, A. Serebrenik, and M. van den Brand. Survival of Eclipse third-party plug-ins. In *Proceedings of the 2012 IEEE International Conference on Software Maintenance*, ICSM'12, pages 368–377, Sept. 2012. 296
207. R. W. Butler and G. B. Finelli. The infeasibility of quantifying the reliability of life-critical real-time software. *IEEE Transactions on Software Engineering*, 19(1):3–12, 1993. 133
208. G. Butts and K. Linton. The joint confidence level paradox: A history of denial. In *NASA 2009 Cost Estimating Symposium*. NASA Center for Aerospace Information, Apr. 2009. 105
209. B. Calder, D. Grunwald, and B. Zorn. Quantifying behavioral differences between C and C++ programs. *Journal of Programming Languages*, 2(4):313–351, 1995. 151
210. J. Calhoun, C. Savoie, M. Randolph-Gips, and I. Bozkurt. Human reliability analysis in spaceflight applications. *Quality and Reliability Engineering International*, 29(6):869–882, Aug. 2013. 42
211. A. Caliskan-Islam, R. Harang, A. Liu, A. Narayanan, C. Voss, F. Yamaguchi, and R. Greenstadt. De-anonymizing programmers via code stylometry. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 255–270, Aug. 2015. 163, 164
212. C. F. Camerer and E. F. Johnson. The process-performance paradox in expert judgment: How can the experts know so much and predict so badly? In K. A. Ericsson and J. Smith, editors, *Towards a general theory of expertise: Prospects and limits*. Cambridge University Press, 1991. 75
213. J. I. D. Campbell. On the relation between skilled performance of simple division and multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(5):1140–1159, 1997. 41
214. M. Campbell-Kelly. *Foundations of Computer Programming in Britain (1945 - 1955)*. PhD thesis, Department of Mathematics and Computer Studies, Sunderland Polytechnic, June 1980. 72
215. M. Campbell-Kelly and D. D. Garcia-Swartz. Economic perspectives on the history of the computer time-sharing industry, 1965–1985. *IEEE Annals of the History of Computing*, 30(1):16–36, Jan.–Mar. 2008. 77
216. M. Caneill and S. Zacchiroli. Debsources: Live and historical views on macro-level software evolution. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM'14, pages 28:1–28:10, Sept. 2014. 87
217. G. Canfora, L. Cerulo, M. Cimitile, and M. Di Penta. Social interactions around cross-system bug fixings: the case of FreeBSD and OpenBSD. In *Proceedings of the 8th Working Conference on Mining Software Repositories*, MSR'11, pages 143–152, May 2011. 214
218. B. Caprile and P. Tonella. Nomen est omen: Analyzing the language of function identifiers. In *Proceedings of the 6<sup>th</sup> Working Conference on Reverse Engineering*, WCRE'99, pages 112–122, Oct. 1999. 161
219. S. Carter-Thomas and E. Rowley-Jolivet. If-conditionals in medical discourse: From theory to disciplinary practice. *Journal of English for Academic Purposes*, 7(3):191–205, July 2008. 37
220. E. Caspi. Empirical study of opportunities for bit-level specialization in word-based programs. Thesis (m.s.), University of California, Berkeley, 2000. 167
221. D. Castelvecchi. The biggest mystery in mathematics: Shinichi Mochizuki and the impenetrable proof. *Nature*, 526(7572):178–181, Oct. 2015. 124
222. J. P. Cavanagh. Relation between the immediate memory span and the memory search rate. *Psychological Review*, 79(6):525–530, 1972. 22
223. M. Ceccato, M. Di Penta, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella. A family of experiments to assess the effectiveness and efficiency of source code obfuscation techniques. *Empirical Software Engineering*, 19(4):1040–1074, 2014. 163
224. C. Cederström and P. Fleming. *Dead Man Working*. Zero books, 2012. 80
225. D. Centola and A. Baronchelli. The spontaneous emergence of conventions: An experimental study of cultural evolution. *PNAS*, 112(7):1989–1994, Feb. 2015. 71
226. D. Centola, R. Willer, and M. Macy. The Emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology*, 110(4):1009–1040, Jan. 2005. 72

227. P. E. Ceruzzi. The early computers of Konrad Zuse, 1935 to 1945. *Annals of the History of Computing*, 3(3):241–262, July 1981. 1
228. F. Chandler, I. A. Heard, M. Presley, A. Burg, E. Midden, and P. Mongan. NASA human error analysis. Technical report, NASA Office of Safety and Mission Assurance, Sept. 2010. 42
229. F. T. Chandler, Y. H. J. Chang, A. Mosleh, J. L. Marble, R. L. Boring, and D. I. Gertman. Human reliability analysis methods: Selection guidance for NASA. Nasa/osma technical report, NASA Headquarters Office of Safety and Mission Assurance, July 2006. 42
230. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection—A survey. *ACM Computing Surveys*, 41(3):1–58, July 2009. 365, 366
231. K. Chandrasekar. *High-Level Power Estimation and Optimization of DRAMs*. PhD thesis, Technische Universiteit Delft, Oct. 2014. 341
232. A. C. Chang and P. Li. Is economics research replicable? Sixty published papers from thirteen journals say “usually not”. *Finance and Economics Discussion Series* 2015-083, Washington: Board of Governors of the Federal Reserve System, Sept. 2015. 7
233. W. Chang. *R Graphics Cookbook*. O'Reilly, 2012. 189
234. A. Chao. Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 45(2):427–438, June 1989. 99
235. A. Chao, R. K. Colwell, C.-W. Lin, and N. J. Gotelli. Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, 90(4):1125–1133, Apr. 2009. 99
236. A. Chao, S.-M. Lee, and S.-L. Jeng. Estimating population size for capture–recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48(1):201–216, Mar. 1992. 99
237. A. Chao and C.-W. Lin. Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics*, 68(3):912–921, Sept. 2012. 99
238. A. Chao and M. C. K. Yang. Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, 80(1):193–201, Mar. 1993. 148
239. G. Charness and U. Gneezy. Strong evidence for gender differences in risk taking. *Journal of Economic Behavior & Organization*, 83(1):50–58, June 2012. 44
240. W. G. Chase and K. A. Ericsson. Skill and working memory. In G. H. Bower, editor, *The Psychology of Learning and Motivation*, pages 1–58. Academic, 1982. 32
241. P. D. Chatzoglou and L. A. Macaulay. Requirements capture and analysis : A survey of current practice. *Requirements Engineering*, 1(2):75–87, June 1996. 117
242. D. D. Chen and G.-J. Ahn. Security analysis of x86 processor microcode. Thesis (B.Sc.), Arizona State University, Dec. 2014. 136
243. T. Chen, Y. Chen, Q. Guo, O. Temam, T. Wu, and W. Hu. Statistical performance comparisons of computers. In *18th International Symposium on High Performance Computer Architecture*, HPCA, pages 1–12, Feb. 2012. 222, 224
244. Y. Chen, A. Groce, X. Fern, C. Zhang, W.-K. Wong, E. Eide, and J. Regehr. Taming compiler fuzzers. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI'13, pages 197–208, June 2013. 146, 276
245. P. W. Cheng, K. J. Holyoak, R. E. Nisbett, and L. M. Oliver. Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, 18(3):293–328, July 1986. 32
246. A. Chesson and G. Chamberlin. Survey-based measures of software investment in the UK. *Economic Trends* 627, Office for National Statistics, UK, Feb. 2006. 70
247. R. N. Chesterman. Report of Queensland health payroll system commission of inquiry. Report, Queensland Government, Australia, July 2013. 105, 110
248. H. Cheung and S. Kemper. Competing complexity metrics and adults’ production of complex sentences. *Applied Psycholinguistics*, 13:53–76, 1992. 155
249. R. C. Cheung. A user-oriented software reliability model. *IEEE Transactions on Software Engineering*, 6(2):118–125, 1980. 213
250. J. Y. Chiao, A. R. Bordeaux, and N. Ambady. Mental representations of social status. *Cognition*, 93(2):B49–B57, Sept. 2004. 41
251. J. J. Chilenski. An investigation of three forms of the modified condition decision coverage (MCDC) criterion. Final Report DOT/FAA/AR-01/18, U.S. Department of Transportation, Federal Aviation Administration, Apr. 2001. 147
252. S. Chilton, J. Covey, M. Jones-Lee, G. Loomes, and H. Metcalf. Valuation of health benefits associated with reductions in air pollution. Technical report, Department for Environment, Food and Rural Affairs, May 2004. 128
253. C.-H. Chiu, Y.-T. Wang, B. A. Walther, and A. Chao. An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. *Biometrics*, 70(3):671–682, Sept. 2014. 99
254. S. Christey and B. Martin. Buying into the bias: Why vulnerability statistics suck. blackhat USA 2013, July-Aug. 2013. 127
255. CRAN task view: Probability distributions. website, June 2016. <http://CRAN.R-project.org/view=Distributions>. 200, 206
256. A. CIA. Analytic thinking and presentation for intelligence producers: Analysis training handbook. Technical report, Office of Training and Education, Central Intelligence Agency, Aug. 1997. 188
257. Z. J. Ciechanowicz and A. C. D. Weever. The ‘completeness’ of the Pascal test suite. *Software—Practice and Experience*, 14(5):463–471, 1984. 148
258. D. Citron. MisSPECulation: Partial and misleading use of SPEC CPU2000 in computer architecture conferences. In *Proceedings of the 30th annual international symposium on Computer architecture*, ISCA'03, pages 52–61, June 2003. 337
259. D. Citron and D. G. Feitelson. "look it up" or "do the math": An energy, area, and timing analysis of instruction reuse and memoization. Technical Report H-0196, IBM, Oct. 2003. 323, 327
260. I. Ciupa, A. Pretschner, M. Oriol, A. Leitner, and B. Meyer. On the number and nature of faults found by random testing. *Software Testing, Verification and Reliability*, 21(1):3–28, Mar. 2011. 146
261. H. H. Clark. *Understanding language*. Cambridge University Press, 1996. 149
262. H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986. 71
263. A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. 276
264. W. S. Cleveland. *The Elements of Graphing Data*. Wadsworth Advanced Book Program, 1985. 189
265. W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, Sept. 1984. 189
266. J. Clune, J.-B. Mouret, and H. Lipson. The evolutionary origins of modularity. *Proceedings of the Royal Society B*, 280(20122863):???, Apr. 2013. 168
267. A. Coad. Investigating the exponential age distribution of firms. *Economics: The Open-Access, Open-Assessment E-Journal*, 4(2010-17):1–30, Mar. 2010. 79
268. N. M. Coe. *The growth and locational dynamics of the UK computer services industry, 1981–1996*. PhD thesis, Department of Geography, University of Durham, 1996. 79
269. J. Coelho and M. T. Valente. Why modern open source projects fail. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE'17, pages 186–196, Sept. 2017. 104
270. J. Cohen. *Statistical Power Analysis for the Behavioural Sciences*. Routledge, second edition, 1988. 232, 233
271. J. Cohen. The Earth is round ( $p < 0.05$ ). *American Psychologist*, 49(12):997–1003, 1994. 228
272. Z. Coker, S. Hasan, J. Overbey, M. Hafiz, and C. Kästner. Integers in C: An open invitation to security attacks? Technical Report CSSE14-01, Auburn University, Feb. 2014. 170
273. M. Cokol, I. Iossifov, R. Rodriguez-Esteban, and A. Rzhetsky. How many scientific papers should be retracted? *European Molecular Biology Organization*, 8(5):422–423, Apr. 2007. 8
274. Numbers every programmer should know. website, Oct. 2016. [https://github.com/colin-scott/interactive\\_latencies](https://github.com/colin-scott/interactive_latencies). 195, 196
275. M. Collard, A. Ruttle, B. Buchanan, and M. J. O'Brien. Population size and cultural evolution in nonindustrial food-producing societies. *PLoS ONE*, 8(9):e72628, Sept. 2013. 71

276. C. Collberg, T. Proebsting, and A. M. Warren. Repeatability and benefit in computer systems research-A study and a modest proposal. Technical Report TR 14-014, Department of Computer Science, University of Arizona, Feb. 2015. 7
277. R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon1, and J. T. Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, Mar. 2012. 99
278. C. Commeyne, A. Abran, and R. Djouab. Effort estimation with story points and cosmic function points - an industry case study. *Software Measurement News*, 21(1):25–36, 2016. 107, 108
279. Committee of Public Accounts. HM revenue and customers: ASPIRE—re-competition of outsourced IT services. Technical Report Twenty-eighth Report of Session 2006-07, UK Parliament, June 2007. 58, 110
280. Comptroller General of the United States. Multiyear leasing and government-wide purchasing of automatic data processing equipment should result in significant savings. Technical Report B-115369, U.S. General Accounting Office, Apr. 1971. 76
281. Comptroller General of the United States. Federal agencies' maintenance of computer programs: Expensive and undermanaged. Technical Report AFMD-81-25, U.S. General Accounting Office, Feb. 1981. 90
282. S. Condon, M. Regardie, M. Stark, and S. Waligora. Cost and schedule estimation study report. Technical Report SEL-93-002, Goddard Space Flight Center, Nov. 1993. 106, 116, 117
283. B. Conrad and M. Mitzenmacher. Power laws for monkeys typing randomly: The case of unequal probabilities. *IEEE Transactions on Information Theory*, 50(7):1403–1414, July 2004. 211
284. J. J. Cook and C. Zilles. A characterization of instruction-level error derating and its implications for error detection. In *IEEE International Conference on Dependable Systems and Networks With FTCS and DCC*, DSN 2008, pages 482–491, June 2008. 242
285. P. Coombs. *IT Project Estimation: A Practical Guide to the Costing of Software*. Cambridge University Press, 2003. 108
286. A. Corazza, V. Maggio, and G. Scanniello. Coherence of comments and method implementations: a dataset and an empirical investigation. *Software Quality Journal*, 26(2):751–777, June 2018. 163
287. J. Corbet, G. Kroah-Hartman, and A. McPherson. Linux kernel development: How fast it is going, who is doing it, what they are doing, and who is sponsoring it? Technical report, The Linux Foundation, Dec. 2010. 103
288. J. Corbet, G. Kroah-Hartman, and A. McPherson. Linux kernel development: How fast it is going, who is doing it, what they are doing, and who is sponsoring it. Technical report, The Linux Foundation, Mar. 2012. 242
289. M. Correll and M. Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, Dec. 2014. 184
290. M. J. Cortese and M. M. Khanna. Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *The Quarterly Journal of Experimental Psychology*, 60(8):1072–1082, Aug. 2007. 161
291. g, a statistical myth. Three-Toed Sloth: blog, Oct. 2007. <http://bactra.org/weblog/523.html>. 43
292. L. Cosmides and J. Tooby. Evolutionary psychology: A primer. Technical report, Center for Evolutionary Psychology, University of California, Santa Barbara, 1998. 16, 36
293. D. L. Costa and M. E. Kahn. Changes in the value of life, 1940-1980. Working Paper No. 9396, National Bureau of Economic Research, USA, Dec. 2002. 128
294. V. Costan and S. Devadas. Intel SGX explained. Technical Report ???, MIT, Jan. 2016. 78
295. D. Cotroneo, R. Pietrantuono, S. Russo, and K. Trivedi. How do bugs surface? A comprehensive study on the characteristics of software bugs manifestation. *The Journal of Systems and Software*, 113(C):27–43, Mar. 2016. 125
296. J. D. Couger and M. A. Colter. *Maintenance Programming: Improving Productivity Through Motivation*. Prentice-Hall, Inc, 1985. 82
297. N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–185, 2001. 21
298. M. F. Cowlishaw. Decimal floating-point: Algorithm for computers. In *Proceedings of the 16th IEEE Symposium on Computer Arithmetic*, pages 104–111, June 2003. 125
299. W. Crooymans, P. Pradhan, and S. Jansen. Exploring network modelling and strategy in the Dutch product software ecosystem. In *Proceedings of the International Conference on Software Business*, page ???, Apr. 2015. 79
300. F. E. Croxton and R. E. Stryker. Bar charts versus circle diagrams. *Journal of the American Statistical Association*, 22(160):473–482, Dec. 1927. 188
301. G. Cumming and R. Maillardet. Confidence intervals and replication: Where will the next mean fall? *Psychological Methods*, 11(3):217–227, 2006. 230
302. C. R. Cummins. *The interpretation and use of numerically-quantified expressions*. PhD thesis, Research Centre for English and Applied Linguistics, University of Cambridge, Nov. 2011. 138
303. B. Curtis, H. Krasner, and N. Iscoe. A field study of the software design process for large systems. *Communications of the ACM*, 31(11):1268–1287, Nov. 1988. 111
304. B. Curtis, S. B. Sheppard, and E. Kruesi. Evaluation of software life cycle data from the PAVE PAWS project. Technical Report RADC-TR-80-28, Rome Air Development Center, Griffiss Air Force Base, Mar. 1980. 114, 115, 126, 136, 138
305. M. A. Cusumano. Factory concepts and practices in software development: An historical overview. Working Paper 3095-89 BPS, Alfred P. Sloan School of Management, Dec. 1989. 121
306. K. Cwalina and B. Abrams. *Framework Design Guidelines: Conventions, Idioms, and Patterns for Reusable .NET Libraries*. Addison-Wesley, 2006. 152
307. J. T. Daly. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Generation Computer Systems*, 22(3):303–312, Feb. 2006. 143
308. A. Damasio. *Self Comes to Mind: Constructing the Conscious Brain*. Vintage books, 2012. 16
309. M. Daneman and P. A. Carpenter. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466, Aug. 1980. 158
310. M. Daneman and P. A. Carpenter. Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(4):561–584, 1983. 158
311. C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, WWW 2013, pages 307–318, May 2013. 160
312. A. Danowitz, K. Kelley, J. Mao, J. P. Stevenson, and M. Horowitz. CPU DB: Recording microprocessor history. *Communications of the ACM*, 55(4):55–63, Apr. 2012. 69, 180, 336
313. J. Darley and C. D. Batson. From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27(1):100–108, 1973. 18
314. P. A. David. Clio and the economics of QWERTY. *The American Economic Review*, 75(2):332–337, May 1985. 68
315. J. W. Davidson, J. R. Rabung, and D. B. Whalley. Relating static and dynamic machine code measurements. Technical Report CS-89-03, Department of Computer Science, University of Virginia, July 1989. 152
316. C. J. Davis. The spatial coding model of visual word identification. *Psychological Review*, 117(3):713–758, July 2010. 22, 149
317. S. J. Davis, J. MacCrisken, and K. M. Murphy. Economic perspectives on software design: PC operating systems and platforms. Working Paper No. 8411, National Bureau of Economic Research, USA, Aug. 2001. 1
318. S. Dayal. Characterizing HEC storage systems at rest. Technical Report CMU-PDL-08-109, Parallel Data Laboratory, Carnegie Mellon University, July 2008. 211
319. R. de Blieck. *Empirical studies on the economic impact of trust*. PhD thesis, Erasmus Research Institute of Management, Rotterdam, May 2015. 64

320. A. D. de Groot. *Thought and Choice in Chess*. Amsterdam University Press, 2008. 31
321. J. L. de la Vara, M. Borg, K. Wnuk, and L. Moonen. An industrial survey of safety evidence change impact analysis practice. *IEEE Transactions on Software Engineering*, 42(12):1095–1117, Dec. 2016. 86, 138
322. B. B. de Mesquita, A. Smith, R. M. Siverson, and J. D. Morrow. *The Logic of Political Survival*. The MIT Press, 2005. 111
323. A. B. de Oliveira, J.-C. Petkovich, T. Reidemeister, and S. Fischmeister. DataMill: Rigorous performance evaluation made easy. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering*, ICPE’13, pages 137–148, Apr. 2013. 344
324. C. B. De Soto, M. London, and S. Handel. Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 2(4):513–521, 1965. 38
325. K. De Vogeleer. *La loi de convexité énergie-fréquence de la consommation des programmes : modélisation, thermosensibilité et applications*. PhD thesis, Informatique [cs] Telecom ParisTech, Sept. 2015. 339
326. G. de Wit. Firm size distributions: An overview of steady-state distributions resulting from firm dynamics models. Technical Report N200418, EIM Business and Policy Research, Jan. 2005. 79
327. I. J. Deary. *Intelligence: A Very Short Introduction*. Oxford University Press, 2001. 43
328. B. K. Debnath, M. F. Mokbel, and D. J. Lilja. Exploiting the impact of database system configuration parameters: A design of experiments approach. *IEEE Data Engineering Bulletin*, 31(1):3–10, Mar. 2008. 327, 328
329. A. Decan, T. Mens, and M. Claes. An empirical comparison of dependency issues in OSS packaging ecosystems. In *IEEE 24th International Conference on Software Analysis, Evolution and Reengineering*, SANER 2017, page ???, Feb. 2017. 92
330. A. Decan, T. Mens, M. Claes, and P. Grosjean. When GitHub meets CRAN: An analysis of inter-repository package dependency problems. In *23rd International Conference on Software Analysis, Evolution, and Reengineering*, SANER’16, page ???, Mar. 2016. 92
331. A. Decan, T. Mens, and E. Constantinou. On the impact of security vulnerabilities in the npm package dependency network. In *15th International Conference on Mining Software Repositories*, MSR 2018, page ???, May 2018. 141
332. S. Dehaene. Symbols and quantities in parietal cortex: elements of a mathematical theory of number representation and manipulation. In P. Haggard, Y. Rossetti, and M. Kawato, editors, *Sensorimotor Foundations of Higher Cognition (Attention and Performance) XXII*, chapter 24, pages 527–574. Oxford University Press, Nov. 2007. 40
333. S. Dehaene. *Reading in the Brain: The Science and evolution of a human invention*. Viking, 2009. 15
334. S. Dehaene. *The Number Sense*. Oxford University Press, revised and updated edition, 2011. 38, 40
335. S. Dehaene, S. Bossini, and P. Giraux. The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, 122(3):371–396, Sept. 1993. 18
336. S. Dehaene, E. Dupoux, and J. Mehler. Is numerical comparison digits? Analogical and symbolic effects in two-digit number comparisons. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3):626–641, 1990. 40
337. S. Dehaene, V. Izard, E. Spelke, and P. Pica. Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigenous cultures. *Science*, 320(5880):1217–1220, May 2008. 17, 40
338. S. M. Dekleva. The influence of the information systems development approach on maintenance. *MIS Quarterly*, 16(3):355–372, Sept. 1992. 86
339. R. T. DeLamarter. *Big Blue: IBM’s Use and Abuse of Power*. Pan Books, 1988. 65, 109
340. S. DellaVigna. Psychology and economics: Evidence from the field. Working Paper No. 13420, National Bureau of Economic Research, USA, Sept. 2007. 49
341. J. Demmel and Y. Hilda. Accurate floating point summation. Technical Report UCB//CSD-02-1180, University of California, Berkeley, May 2002. 123
342. Department of Defense. Military standard DOD-STD-2167 defense system software development. Standard DOD-STD-2167, US Department of Defense, 1985. 113
343. Department of Defense. Standard practice system safety. Standard MIL-STD-882E, U.S. Department of Defense, May 2012. 128
344. G. Destefanis. Which programming language should a company use? A Twitter-based analysis. Technical Report CRIM-14/10-23-MODL, Computer Research Institute of Montréal, Oct. 2014. 90
345. S. Deutsch and M. H. Jørgensen. Studying the hidden costs of offshoring – the effect of psychic distance. Thesis (m.s.), Copenhagen Business School, Aug. 2014. 108
346. J. P. DeVale. *High Performance Robust Computer Systems*. PhD thesis, Electrical and Computer Engineering, Pittsburgh, Oct. 2001. 131
347. A. Di Franco, H. Guo, and C. Rubio-González. A comprehensive study of real-world numerical bug characteristics. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, ASE’2017, pages 509–519, Nov. 2017. 136
348. C. Di Martino, Z. Kalbarczyk, R. K. Iyer, J. F. Fabio Baccanico and, and W. Kramer. Lessons learned from the analysis of system failures at petascale: The case of Blue Waters. In *44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, DSN 2014, pages 610–621, June 2014. 142
349. M. Di Penta, L. Cerulo, and L. Aversano. The life and death of statistically detected vulnerabilities: an empirical study. *Information and Software Technology*, 51(10):1469–1484, Oct. 2009. 129, 303
350. T. F. Dickey. Programmer variability. *Proceedings of the IEEE*, 69(7):844–845, July 1981. 74
351. L. S. Dickstein. The effect of figure on syllogistic reasoning. *Memory & Cognition*, 6(1):76–83, 1978. 37
352. A. Diekmann. Not the first digit! Using Benford’s law to detect fraudulent scientific data. *Journal of Applied Statistics*, 34(3):321–329, Oct. 2007. 372
353. J. Dietrich, K. Jezek, and P. Brada. What Java developers know about compatibility, and why this matters. In *eprint arXiv:cs.SE/1408.2607v1*, Aug. 2014. 346
354. C. DiMarco, G. Hirst, and M. Stede. The semantic and stylistic differentiation of synonyms and near-synonyms. In *AAAI Spring Symposium on Building Lexicons for Machine Translation*, pages 114–121, Mar. 1993. 199
355. A. Dinaburg. Bitsquatting: DNS hijacking without exploitation. Reference 2011-307, Raytheon Company, July 2011. 143
356. D. K. Dirlam. Most efficient chunk sizes. *Cognitive Psychology*, 3(2):355–359, Apr. 1972. 24
357. H. Do, S. Mirarab, L. Tahvildari, and G. Rothermel. An empirical study of the effect of time constraints on the cost-benefits of regression testing. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2008, pages 71–82, Nov. 2008. 148
358. C. Domas. Breaking the x86 ISA. blackhat USA 2017, July 2017. 136
359. D. J. Dooling and R. E. Christiaansen. Episodic and semantic aspects of memory for prose. *Journal of Experimental Psychology: Human Learning and Memory*, 3(4):428–436, 1977. 157
360. J. R. Douceur and W. J. Bolosky. A large-scale study of file-system contents. In *Proceedings of the 1999 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*, SIGMETRICS’99, pages 59–70, July 1999. 211
361. J. R. Doyle. Survey of time preference, delay discounting models. *Judgment and Decision Making*, 8(2):116–135, Mar. 2013. 49
362. G. Dréan. *The Computer Industry: Structure, economics, perspectives*. Gérard Dréan, english edition, 2012. 78
363. S. Drobisz, T. Mens, and R. Di Cosmo. A historical analysis of Debian package conflicts. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, MSR’15, pages 212–223, June 2015. 92
364. R. I. M. Dunbar and R. Sosis. Optimising human community sizes. *Evolution and Human Behavior*, 39(1):106–111, Jan. 2018. 70
365. J. . R. Dunham and L. A. Lauterbach. An experiment in software reliability additional analyses using data from automated replications. NASA Contractor Report 178395, Research Triangle Institute, North Carolina, Jan. 1988. 132

366. J. R. Dunham and J. L. Pierce. An experiment in software reliability. NASA Contractor Report 172553, NASA Langley Research Center, Mar. 1986. 132, 194
367. L. M. Dunn. *An Investigation of the Factors Affecting the Lifecycle Costs of COTS-Based Systems*. PhD thesis, School of Computing, University of Portsmouth, England, June 2011. 85
368. D. Dunning, C. Heath, and J. M. Suls. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5(3):69–106, Apr. 2004. 346
369. T. Dybå, V. B. Kampenes, and D. I. K. Sjøberg. A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, 48(8):745–755, Aug. 2006. 5, 318
370. R. Dyer, H. Rajan, H. A. Nguyen, and T. N. Nguyen. Mining billions of AST nodes to study actual and potential usage of Java language features. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE’14, pages 779–790, May-June 2014. 167
371. A. Eckbreth, C. Saff, K. Connolly, N. Crawford, C. Eick, M. Goorsky, N. Kacena, D. Miller, R. Schafrik, D. Schmidt, D. Stein, M. Stroscio, G. Washington, and J. Zolper. Sustaining Air Force aging aircraft into the 21<sup>st</sup> century. Technical Report SAB-TR-11-01, United States Air Force Scientific Advisory Board, Aug. 2011. 84
372. The changing US technology sector. Daily chart for April 21 2015 on The Economist webpage, Apr. 2015. As of Q1 2015, Sources: Thomson Reuters; awk scripts+R converted the data embedded in Javascript. 2, 3
373. EDB. Offensive security’s exploit database archive. <https://www.exploit-db.com/>, Mar. 2018. 126
374. S. Eder, M. Junker, E. Jürgens, B. Hauptmann, R. Vaas, and K.-H. Prommer. How much does unused code matter for maintenance? In *34th International Conference on Software Engineering*, ICSE, pages 1102–1111, June 2012. 57, 93
375. A. Edmundson, B. Holtkamp, E. Rivera, M. Finifter, A. Mettler, and D. Wagner. An empirical study on the effectiveness of security code review. In *Proceedings of the 5th international conference on Engineering Secure Software and Systems*, ESSoS’13, pages 197–212, Feb. 2013. 230, 248, 258
376. M. A. Edwards and S. Roy. Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, 34(1):51–61, Jan. 2017. 6
377. K. Ehrlich and P. N. Johnson-Laird. Spatial descriptions and referential continuity. *Journal of Verbal Learning and Verbal Behavior*, 21(3):296–306, June 1982. 159
378. S. G. Eick, C. R. Loader, M. D. Long, L. G. Votta, and S. V. Wiel. Estimating software fault content before coding. In *Proceedings of the 14th international conference on Software engineering*, ICSE’92, pages 59–65, May 1992. 144
379. P. Ein-Dor. Grosch’s law re-revisited: CPU power and the cost of computation. *Communications of the ACM*, 28(2):142–151, Feb. 1985. 78
380. T. Eisensee and D. Strömberg. News droughts, news floods, and U.S. disaster relief. *The Quarterly Journal of Economics*, 122(2):693–728, May 2007. 127
381. K. El Emam, S. Benlarbi, N. Goel, W. Melo, H. Lounis, and S. N. Rai. The optimal class size for object-oriented software. *IEEE Transactions on Software Engineering*, 28(5):494–509, Mar. 2002. 194
382. K. El Emam and A. G. Koru. A replicated survey of IT software project failures. *IEEE Software*, 25(5):84–90, Apr. 2008. 104
383. A. Elci. The dependence of operating system size upon allocatable resources. Technical Report 75-172, Department of Computer Science, Purdue University, Dec. 1975. 70
384. J. Elliott, M. Hoemmen, and F. Mueller. Exploiting data representation for fault tolerance. In *eprint arXiv:cs.NA/1312.2333v1*, Dec. 2013. 123
385. N. C. Ellis and R. A. Hennelly. A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology*, 71:43–51, 1980. 21, 322
386. P. D. Ellis. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010. 233
387. R. Engbert, A. Nuthmann, E. M. Richter, and R. Kliegl. SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777–813, Apr. 2005. 34
388. J. Engblom. Why SpecInt95 should not be used to benchmark embedded systems tools. *ACM SIGPLAN Notices*, 34(7):96–103, July 1999. 74, 165
389. J. Ensign and D. K. Akaka. Defense acquisitions: DOD has paid billions in award and incentive fees regardless of acquisition outcomes. Technical Report GAO-06-66, United States Government Accountability Office, Dec. 2005. 111
390. Y.-H. Eom and H.-H. Jo. Generalized friendship paradox in complex networks: The case of scientific collaboration. In *eprint arXiv:cs.SI/1401.1458*, Apr. 2014. 68
391. D. M. Erceg-Hurn and V. M. Mirosevich. Modern robust statistical methods. *American Psychologist*, 63(7):591–601, Oct. 2008. 219
392. K. A. Ericsson and N. Charness. Expert performance. *American Psychologist*, 49(8):725–747, Aug. 1994. 31, 75
393. K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3):363–406, 1993. also University of Colorado, Technical Report #91-06. 31, 75
394. K. A. Ericsson and A. C. Lehmann. Expert and exceptional performance: Evidence of maximal adaption to task constraints. *Annual Review of Psychology*, 47:273–305, 1996. 31
395. K. Eriksson, D. H. Bailey, and D. C. Geary. The grammar of approximating number pairs. *Memory & Cognition*, 38(3):333–343, Apr. 2010. 137
396. L. Eshkevari, F. D. Santos, J. R. Cordy, and G. Antoniol. Are PHP applications ready for hack? In *IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering*, SANER 2015, pages 63–72, Mar. 2015. 170
397. L. M. Eshkevari, V. Arnaoudova, M. Di Penta, R. Oliveto, Y.-G. Guéhéneuc, and G. Antoniol. An exploratory study of identifier renamings. In *Proceedings of the 8th Working Conference on Mining Software Repositories*, MSR’11, pages 33–42, May 2011. 95
398. H. Esmaeilzadeh, T. Cao, X. Yang, S. M. Blackburn, and K. S. McKinley. Looking back on the language and hardware revolutions: Measured power, performance, and scaling. In *Proceedings of the sixteenth international conference on Architectural support for programming languages and operating systems*, ASPLOS XVI, pages 319–332, Mar. 2011. 225
399. W. K. Estes. *Classification and Cognition*. Oxford University Press, 1994. 29
400. J. S. B. T. Evans, J. L. Barston, and P. Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11(3):295–306, 1983. 37
401. J. L. Eveleens and C. Verhoef. The rise and fall of the Chaos report figures. *IEEE Software*, 27(1):30–36, Jan. 2010. 104
402. J. Eyolfson, L. Tan, and P. Lam. Do time of day and developer experience affect commit bugginess? In *Proceedings of the 8th Working Conference on Mining Software Repositories*, MSR’11, pages 153–162, May 2011. 103, 284, 285, 304
403. J. Eyolfson, L. Tan, and P. Lam. Correlations between bugginess and time-based commit characteristics. *Empirical Software Engineering*, 19(4):1009–1039, Aug. 2014. 305, 306, 307
404. Facebook. Facebook Inc. 2013 Form 10-K. website: accessed on 25 Feb 2017, 2014. <https://www.sec.gov/Archives/edgar/data/1326801/000132680114000007/fb-12312013x10k.htm>. 64
405. Facebook. Facebook Inc. 2015 Form 10-K. website: accessed on 13 Feb 2017, 2016. <https://www.sec.gov/Archives/edgar/data/1326801/000132680116000043/fb-12312015x10k.htm>. 64
406. R. Falk and C. Konold. Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104(2):301–318, 1997. 137
407. D. Fanelli. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, 4(5):e5738, May 2009. 8
408. D. Fanelli. "Positive" results increase down the hierarchy of the sciences. *PLoS ONE*, 5(4):e10068, Apr. 2010. 7
409. F. C. Fang, R. G. Steen, and A. Casadevall. Misconduct accounts for the majority of retracted scientific papers. *PNAS*, 109(42):17028–17033, Oct. 2012. 8

410. M. Fang and M. Hafiz. Discovering buffer overflow vulnerabilities in the wild: An empirical study. In *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM'14, pages 23:1–23:10, Sept. 2014. [127](#)
411. L. Farr and B. Nanus. Factors that affect the cost of computer programming, volume I. Technical Documentary Report ESD-TDR-64-448, United States Air Force, L. G. Hanscom Field, Bedford, Massachusetts, July 1964. [106](#)
412. L. Farr and H. J. Zagorski. Factors that affect the cost of computer programming, volume II: A quantitative analysis. Technical Documentary Report ESD-TDR-64-448, United States Air Force, L. G. Hanscom Field, Bedford, Massachusetts, Sept. 1964. [106](#)
413. J. Farrell and P. Klempner. Coordination and lock-in: Competition with switching costs and network effects. In M. Armstrong and R. H. Porter, editors, *Handbook of Industrial Organization, Volume 3*, chapter 31, pages 1967–2072. North-Holland, Oct. 2007. [58](#)
414. J. Farrell and C. Shapiro. Dynamic competition with switching costs. *RAND Journal of Economics*, 19(1):123–137, 1988. [58](#)
415. FDA. General principles of software validation. Final guidance for industry and fda staff, U.S. Food and Drug Administration, Jan. 2002. [128](#)
416. Federal Trade Commission. Dell computer corporation consent order, etc., in regard to alleged violation of sec. 5 of the federal trade commission act, docket c-3658. In P. C. Epperson, editor, *Federal Trade Commission decisions: Findings, opinions and orders volume 121*, pages 616–643. U.S. Government Printing Office, May 1996. [80](#)
417. D. G. Feitelson. *Workload Modeling for Computer Systems Performance Evaluation*. Cambridge University Press, 2014. [77](#), [337](#), [372](#)
418. D. G. Feitelson and B. Nitzberg. Job characteristics of a production parallel scientific workload on the NASA Ames iPSC/860. In D. G. Feitelson and L. Rudolph, editors, *Job Scheduling Strategies for Parallel Processing: Lecture Notes Computer Science vol. 949*, chapter 19, pages 337–360. Springer-Verlag, June 1995. [363](#)
419. S. L. Feld. Why your friends have more friends than you do. *The American Journal of Sociology*, 96(6):1464–1477, May 1991. [68](#)
420. J. Feldman. Minimization of boolean complexity in human concept learning. *Nature*, 407:630–633, Oct. 2000. [29](#)
421. J. Feldman. An algebra of human concept learning. *Journal of Mathematical Psychology*, 50(4):339–368, Aug. 2006. [29](#)
422. M. Felici. *Observational Models of Requirements Evolution*. PhD thesis, School of Informatics University of Edinburgh, 2004. [117](#)
423. S. Feng, S. Gupta, A. Ansari, and S. Mahlke. Shoestring: Probabilistic soft error reliability on the cheap. In *Proceedings of the fifteenth edition of ASPLOS on Architectural support for programming languages and operating systems*, ASPLOS'10, pages 385–396, Mar. 2010. [142](#)
424. N. Fenton, M. Neil, W. Marsh, P. Hearty, Ł. Radliński, and P. Krause. On the effectiveness of early life cycle defect prediction with Bayesian nets. *Empirical Software Engineering*, 13(5):499–537, Oct. 2008. [249](#), [250](#)
425. D. V. Ferens and D. S. Christensen. Calibrating software cost models to Department of Defense databases-A review of ten studies. *ISPA Journal of Parametrics*, XVIII(2):55–74, Nov. 1998. [106](#)
426. C. J. Ferguson and M. Heene. A vast graveyard of undead theories publication bias and psychological science's aversion to the Null. *Perspectives on Psychological Science*, 7(6):555–561, Nov. 2012. [7](#)
427. P. Fernández. Valuing real options: Frequently made errors. SSRN Working Paper n. 274855, Instituto de Estudios Superiores de la Empresa, Madrid, June 2001. [56](#)
428. L. Ferrand, M. Brysbaert, E. Keuleers, B. New, P. Bonin, A. Méot, M. Augustinova, and C. Pallier. Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from Chronolex. *frontiers in Psychology*, 2(306), Nov. 2011. [161](#)
429. S. Ferson, J. O'Rawe, A. Antonenko, J. Siegrist, J. Mickley, C. C. Luhmann, K. Sentz, and A. M. Finkel. Natural language of uncertainty: numeric hedge words. *International Journal of Approximate Reasoning*, 57:19–39, Feb. 2015. [138](#)
430. R. G. Fichman and C. F. Kemerer. Incentive compatibility and systematic software reuse. *Journal of Systems and Software*, 57(1):45–60, Apr. 2001. [95](#)
431. A. Filippin and P. Crosetto. A reconsideration of gender differences in risk attitudes. IZA DP No. 8184, The Institute for the Study of Labor, Bonn, May 2014. [44](#)
432. C. J. Fillmore. Topics in lexical semantics. In R. W. Cole, editor, *Current Issues in Linguistic Theory*, pages 76–138. Indiana University Press, 1977. [31](#)
433. Financial Accounting Standards Board. Statement of financial accounting standards no. 86. Technical report, Financial Accounting Foundation, Aug. 1985. [60](#)
434. M. Finifter. Towards evidence-based assessment of factors contributing to the introduction and detection of software vulnerabilities. Technical Report UCB/EECS-2013-49, Electrical Engineering and Computer Sciences, University of California at Berkeley, May 2013. [144](#)
435. M. Finifter, D. Akhawe, and D. Wagner. An empirical study of vulnerability rewards programs. In *Proceedings of the 22nd USENIX conference on Security*, SEC'13, pages 273–288, Aug. 2013. [129](#)
436. E. Fischer. The evolution of character codes, 1874–1968. Nov. 2002. [72](#)
437. D. A. Fisher. A common programming language for the Department of Defense – background and technical requirements. PAPER P-1191, Institute for Defense Analyses, Science and Technology Division, June 1976. [89](#)
438. J. Fisher and R. A. Hinde. The opening of milk bottles by birds. *British Birds*, 42(11):347–357, 1949. [71](#)
439. J. C. Fisher and R. H. Pry. A simple substitution model of technological change. *Technological Forecasting and Social Change*, 3:75–88, Apr. 1971–1972. [62](#)
440. K. Flamm. *Targeting the Computer*. The Brookings Institution, Washington, D.C., 1987. [5](#), [76](#)
441. K. Flamm. *Creating the Computer*. The Brookings Institution, Washington, D.C., 1988. [1](#)
442. K. Flamm. Measuring Moore's law: Evidence from price, cost, and quality indexes. Working Paper No. 24553, National Bureau of Economic Research, Apr. 2018. [4](#)
443. D. Flater. Estimation of uncertainty in application profiles. NIST TN.1826, National Institute of Standards and Technology, Apr. 2014. [344](#)
444. D. Flater. Screening for factors affecting application performance in profiling measurements. NIST Technical Note 1855, National Institute of Standards and Technology, Oct. 2014. [342](#)
445. D. Flater and W. F. Guthrie. A case study of performance degradation attributable to run-time bounds checks on C++ vector access. *Journal of Research of the National Institute of Standards and Technology*, 118(012):260–279, May 2013. [167](#), [252](#), [253](#)
446. P. J. Fleming and J. J. Wallace. How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221, Mar. 1986. [332](#)
447. J. I. Flombaum, J. A. Junge, and M. D. Hauser. Rhesus monkeys (*macaca mulatta*) spontaneously compute addition operations over large numbers. *Cognition*, 97(3):315–325, Oct. 2005. [40](#)
448. B. Floyd, T. Santander, and W. Weimer. Decoding the representation of code in the brain: An fMRI study of code review and expertise. In *Proceedings of the 39th International Conference on Software Engineering*, ICSE'17, pages 175–186, May 2017. [149](#)
449. B. Flyvbjerg. How planners deal with uncomfortable knowledge: The dubious ethics of the American Planning Association. *Cities*, 32:157–163, June 2013. [105](#)
450. B. Flyvbjerg, M. S. Holm, and S. L. Buhl. Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, 68(3):279–295, June 2002. [105](#)
451. J. Fodor. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, 1983. [16](#)
452. R. A. Foley. An evolutionary and chronological framework for human social behaviour. *Proceedings of the British Academy*, 88:95–117, 1996. [15](#)
453. P. Fonseca, K. Zhang, X. Wang, and A. Krishnamurthy. An empirical study on the correctness of formally verified distributed systems. In *Proceedings of the Twelfth European Conference on Computer Systems*, Eurosyst'17, pages 328–343, Apr. 2017. [124](#)
454. C. E. Ford and S. A. Thompson. Conditionals in discourse: A text-based study from English. In E. C. Traugott, A. T. Meulen, J. S. Reilly, and C. A. Furgason, editors, *On Conditionals*, chapter 18, pages 353–372. Cambridge University Press, 1986. [37](#)

455. J. Förster, E. Higgins, and A. T. Bianco. Speed/accuracy decisions in task performance: Built-in trade-off or separate strategic concerns? *Organizational Behavior and Human Decision Processes*, 90(1):148–164, Jan. 2003. 19
456. M. Fowler, K. Beck, J. Brant, W. Opdyke, and D. Roberts. *Refactoring: Improving the Design of Existing Code*. Addison-Wesley, 1999. 6
457. W. B. Frakes, C. J. Fox, and B. A. Nejmeh. *Software Engineering in the Unix/C Environment*. Prentice-Hall, Inc, 1991. 152
458. S. Frederick, G. Loewenstein, and T. O'Donoghue. Time discounting: A critical review. *Journal of Economic Literature*, 40(2):351–401, June 2002. 150
459. D. P. Freedman and G. M. Weinberg. *Handbook of Walkthroughs, Inspections, and Technical Reviews*. Dorset House Publishing, 1990. 145
460. P. A. Freund and N. Kasten. How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, 138(2):296–321, Mar. 2011. 17
461. W.-T. Fu and W. D. Gray. Memory versus perceptual-motor trade-offs in a blocks world task. In *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*, pages 154–159. Erlbaum, 2000. 42
462. Y. Funami and M. H. Halstead. A software physics analysis of akiyama's debugging data. Technical Report CSD-TR 144, Purdue University, May 1975. 164
463. B. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):1–53, June 2010. 364
464. C. A. Furia. Bayesian statistics in software engineering: Practical guide and case studies. In *eprint arXiv:cs.SE/1608.06865*, Aug. 2016. 114, 221
465. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. Statistical semantics: Analysis of the potential performance of keyword information systems. *The Bell System Technical Journal*, 62(6):1753–1805, July-Aug. 1983. 73
466. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, Nov. 1987. 73
467. G. Fursin, R. Miceli, A. Lokhmotov, M. Gerndt, M. Baboulin, A. D. Malony, Z. Chamski, D. Novillo, and D. D. Vento. Collective mind: Towards practical and collaborative auto-tuning. *Scientific Programming Journal*, 22(4):309–329, July 2014. 89
468. T. Futagami, M. Itoh, Y. Mihara, F. Mitsuhashi, H. Nishiyama, M. Shukuguchi, N. Tachi, K. Toyama, H. Obata, Y. Ooizumi, T. Shimizu, and S. Takeichi. *ESCR Embedded System development Coding Reference guide [C Language Edition]*. Information-technology Promotion Agency, Japan, 2.0 edition, 2017. 152
469. R. Futrell, K. Mahowald, and E. Gibson. Large-scale evidence of dependency length minimization in 37 languages. *PNAS*, 112(33):10336–10341, Aug. 2015. 155
470. M. T. Gailliot and R. F. Baumeister. The physiology of willpower: Linking blood glucose to self-control. *Personality and Social Psychology Review*, 11(4):303–327, Nov. 2007. 50
471. W. A. Gale. Good-Turing smoothing without tears. Technical Report 94.5, AT&T Bell Laboratories, Aug. 1994. 369
472. K. Gallaba, A. Mesbah, and I. Beschastnikh. Don't call us, we'll call you: Characterizing callbacks in JavaScript. In *International Symposium on Empirical Software Engineering and Measurement*, ESEM'15, pages 247–256, Oct. 2015. 170, 171
473. T. J. Gandomani, K. T. Wei, and A. K. Binhamid. A case study research on software cost estimation using experts' estimates, Wideband Delphi, and Planning Poker technique. *International Journal of Software Engineering and Its Applications*, 8(11):173–182, Apr. 2014. 329, 330
474. A. Gandy. *The entry of established electronics companies into the early computer industry in the UK and USA*. PhD thesis, London School of Economics and Political Science, 1992. 1, 70
475. J. D. Gannon. An experimental evaluation of data type conversions. *Communications of the ACM*, 20(8):584–595, Aug. 1977. 153
476. Z. Gao, Y. Liang, M. B. Cohen, A. M. Memon, and Z. Wang. Making system user interactive tests repeatable: When and what should we control? In *Proceedings of the 37th International Conference on Software Engineering*, ICSE'15, pages 55–65, June 2015. 147
477. M. K. Gardner, E. Z. Rothkopf, R. Lapan, and T. Laferty. The word frequency effect in lexical decision: Finding a frequency-based component. *Memory & Cognition*, 15(1):24–28, 1987. 162
478. M. R. Garman. The generalizability of private sector research on software project management in two USAF organizations: An exploratory study. Thesis (m.s.), Air Force Institute of Technology, U.S.A., Mar. 2003. 102
479. R. Garner and F. R. Dill. The legendary IBM 1401 data processing system. *IEEE Solid-State Circuits Magazine*, 2(1):28–39, Jan. 2010. 76
480. Gartner. Worldwide smartphone sales. [https://en.wikipedia.org/wiki/Mobile\\_operating\\_system](https://en.wikipedia.org/wiki/Mobile_operating_system), July 2017. 3, 68
481. D. C. Gause and G. M. Weinberg. *Exploring Requirements: Quality before design*. Dorset House Publishing, 1989. 113
482. J. E. Gayek, L. G. Long, K. D. Bell, R. M. Hsu, and R. K. Larson. Software cost and productivity model. Technical Report ATR-2004(8311)-1, Aerospace Corporation, Feb. 2004. 60
483. Y. Ge and B. Xu. Dynamic staffing and rescheduling in software project management: A hybrid approach. *PLoS ONE*, 11(6):e0157104, June 2016. 112
484. Y. Geffen and S. Maoz. On method ordering. In *IEEE 24th International Conference on Program Comprehension*, ICPC'16, page ???, May 2016. 171
485. W. Gellerich, M. Kosiol, and E. Ploedereder. Where does GOTO go to? In *Reliable Software Technology – Ada-Europe 1996*, volume 1088 of *LNCS*, pages 385–395. Springer, 1996. 169
486. S. A. Gelman and E. M. Markman. Categories and induction in young children. *Cognition*, 23:183–209, 1986. 27
487. D. Gentner and S. Goldin-Meadow. *Language In Mind: Advances in the Study of Language and Thought*. MIT Press, 2003. 30, 166
488. S. L. Gerhart and L. Yelowitz. Observations of fallibility in applications of modern programming methodologies. *IEEE Transactions on Software Engineering*, SE-2(3):195–207, Sept. 1976. 124
489. D. M. German, Y. Manabe, and K. Inoue. A sentence-matching method for automatic license identification of source code files. In *Proceedings of the IEEE/ACM international conference on Automated software engineering*, ASE'10, pages 437–446, Apr. 2010. 92
490. E. H. Gibbs, G. A. Munroe, A. M. Zeman, and C. T. Cottingham. JANET SKOLD and DAVID DOSSANTOS, on behalf of themselves and all others similarly situated and the general public, v. INTEL CORPORATION, HEWLETT PACKARD COMPANY and DOES 1-50, case no. 1-05-CV-039231, filing #g-43414. Opinion, Superior court of the state of California for the county of Santa Clara, 2012. 336
491. E. Gibson. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68(1):1–76, Aug. 1998. 155
492. E. Gibson and J. Thomas. Memory limitations and structured forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248, 1999. 159
493. G. Gigerenzer. Striking a blow for sanity in theories of rationality. In M. Augier and J. G. March, editors, *Models of a man: Essays in memory of Herbert A. Simon*, pages 389–409. MIT Press, May 2004. 44
494. G. Gigerenzer. *Rationality for Mortals-How People cope with Uncertainty*. Oxford University Press, 2008. 36, 227
495. G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2):53–96, Apr. 2008. 196
496. G. Gigerenzer, S. Krauss, and O. Vitouch. The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan, editor, *The Sage handbook of quantitative methodology for the social sciences*, chapter 21, pages 391–408. SAGE Publications, Inc, 2004. 228
497. G. Gigerenzer, P. M. Todd, and The ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, 1999. 16, 36, 44

498. B. Gilchrist and R. E. Weber. Employment of trained computer personnel—A quantitative survey. In *Proceedings of the Spring Joint Computer Conference*, AFIPS'72, pages 641–648, May 1972. 81
499. J. Gimpel. Software that checks software: The impact of PC-lint. *IEEE Software*, 31(1):15–19, Jan.-Feb. 2014. 120
500. V. Girotto, A. Mazzocco, and A. Tasso. The effect of premise order on conditional reasoning: a test of the mental model theory. *Cognition*, 63:1–28, 1997. 38
501. M. Givon, V. Mahajan, and E. Muller. Software piracy: Estimation of lost sales and the impact on software diffusion. *Journal of Marketing*, 59(1):29–37, Jan. 1995. 63, 291
502. T. J. Gauthier. Computer time sharing: Its origins and development. *Computers and Automation*, 16(10):23–27, Oct. 1967. 1
503. A. Glenberg. Few believe the world is flat: How embodiment is changing the scientific understanding of cognition. *Canadian Journal of Experimental Psychology*, 69(2):165–171, June 2015. 17
504. F. Gobet. *Understanding Expertise: A Multi-disciplinary Approach*. Palgrave, 2016. 31
505. D. R. Godden and A. D. Baddeley. Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, 66(3):325–331, 1975. 23
506. M. W. Godfrey and Q. Tu. Evolution in open source software: A case study. In *16<sup>th</sup> International Conference on Software Maintenance*, pages 131–142, Oct. 2000. 247
507. A. L. Goel. An experimental investigation into software reliability. Final Technical Report RADC-TR-88-213, CASE Center, Syracuse University, Oct. 1988. 139
508. S. S. Gokhale and R. E. Mullen. The marginal value of increased testing: An empirical analysis using four code coverage measures. *Journal of the Brazilian Computer Society*, 12(3):13–30, Dec. 2006. 148
509. M. M. Gold. A methodology for evaluating time-shared computer system usage. Technical report, Carnegie Mellon University, Aug. 1967. 88
510. L. R. Goldberg, J. A. Johnson, H. W. Eber, R. Hogan, M. C. Ashton, C. R. Cloninger, and H. G. Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006. 43
511. M. S. Goldberg and A. Touw. Statistical methods for learning curves and cost analysis. Technical Report CIMD0006870.A3/1Rev, The CNA Corporation, Mar. 2003. 76
512. E. Goldvarg and P. Johnson-Laird. Naive causality: a mental model theory of causal meaning and reasoning. *Cognitive Science*, 25(4):565–610, July 2001. 39
513. P. Golle. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on Privacy in electronic society*, WPES'06, pages 77–80, Oct. 2006. 363
514. J. M. González-Barahona, G. Robles, I. Herráiz, and F. Ortega. Studying the laws of software evolution in a long-lived FLOSS project. *Journal of Software: Evolution and Process*, 26(7):589–612, July 2014. 183, 223, 276, 292
515. B. H. Good, Y.-A. de Montjoye, and A. Clauset. The performance of modularity maximization in practical contexts. In *eprint arXiv:physics.data-an/0910.0165v2*, Apr. 2010. 214
516. J. Goodman. Lessons learned from seven Space Shuttle missions. NASA Contractor Report CR-2007-213697, Lyndon B. Johnson Space Center, Jan. 2007. 124
517. P. Goodridge, J. Haskel, and G. Wallis. Estimating UK investment in intangible assets and intellectual property rights. Technical Report No. 2014/36, Intellectual Property Office, UK government, Sept. 2014. 4, 77
518. Google books ngram dataset. website, 2015. <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>. 371
519. A. Gopal and B. R. Koka. The role of contracts on quality and returns to quality in offshore software development outsourcing. *Decision Sciences*, 41(3):491–516, Aug. 2010. 110
520. R. Gopinath, A. Alipour, I. Ahmed, C. Jensen, and A. Groce. How hard does mutation analysis have to be, anyway? In *IEEE 26th International Symposium on Software Reliability Engineering*, ISSRE 2015, pages 216–227, Nov. 2015. 148, 199
521. R. Gopinath, C. Jensen, and A. Groce. Code coverage for suite evaluation by developers. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE'14, pages 72–82, June 2014. 147, 148, 263
522. R. Gopinath, C. Jensen, and A. Groce. Mutations: How close are they to real faults? In *IEEE 25th International Symposium on Software Reliability Engineering*, ISSRE'14, pages 189–200, Nov. 2014. 140, 148
523. R. D. Gordon and M. H. Halstead. An experiment comparing Fortran programming times with the software physics hypothesis. Technical Report TR 167, Purdue University, Oct. 1975. 164
524. R. J. Gordon. The postwar evolution of computer prices. Working Paper No. 2227, National Bureau of Economic Research, Apr. 1987. 3
525. N. J. Gotelli and A. Chao. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In S. A. Levin, editor, *Encyclopedia of Biodiversity: vol 5*, pages 195–211. Academic Press, second edition, Feb. 2013. 97
526. M. Gottscho. ViPZonE: Exploiting DRAM power variability for energy savings in Linux x86-64. Thesis (m.s.), Electrical Engineering, UCLA, Mar. 2014. 341
527. M. Gottscho, A. A. Kagalwalla, and P. Gupta. Power variability in contemporary DRAMs. *IEEE Embedded Systems Letters*, 4(12):37–40, June 2012. 341
528. S. Götz, T. Ilsche, J. Cardoso, J. Spillner, U. Aßmann, W. Nagel, and A. Schill. Energy-efficient data processing at sweet spot frequencies. In *OTM 2014 Workshops*, pages 154–171, Apr. 2014. 338
529. G. Gousios and A. Zaidman. A dataset for pull-based development research. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR 2014, pages 368–371, May 2014. 179
530. G. Gousios, A. Zaidman, M.-A. Storey, and A. van Deursen. Work practices and challenges in pull-based development: The integrator's perspective. In *Proceedings of the 37th International Conference on Software Engineering*, pages 358–368, May 2015. 185
531. K. Goševa-Popstojanova, M. Hamill, and R. Perugupalli. Large empirical case study of architecture-based software reliability. In *Proceedings of the 16th IEEE International Symposium on Software Reliability Engineering*, ISSRE'05, pages 43–52, Nov. 2005. 212
532. E. M. Grabbe, S. Ramo, and D. E. Wooldridge. *Handbook of Automation, Computation, and Control, Volume 2: Computers and Data Processing*. John Wiley & Sons, Inc, 1959. 89
533. P. Grady. *Termination of the SIREN ICT project*. Grant Thornton UK LLP, June 2014. 114
534. R. B. Grady and D. L. Caswell. *Software Metrics: Establishing a company-wide program*. Prentice-Hall, Inc, 1987. 126
535. A. C. Graesser, N. L. Hoffman, and L. F. Clark. Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior*, 19(2):135–151, Apr. 1980. 159
536. A. C. Graesser, S. B. Woll, D. J. Kowalski, and D. A. Smith. Memory for typical and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):503–515, June 1980. 157
537. S. Graillat, F. Jézéquel, R. Picot, F. Févotte, and B. Lathuilière. Auto-tuning for floating-point precision with discrete stochastic arithmetic. HAL Id: hal-01331917, HAL archives-ouvertes.fr, June 2016. 125
538. E. E. Grant and H. Sackman. An exploratory investigation of programmer performance under on-line and off-line conditions. *IEEE Transactions on Human Factors in Electronics*, 8(1):33–48, Mar. 1967. 50, 74, 331
539. Graphviz-graph visualization software. website, 2015. <http://www.graphviz.org>. 185
540. C. A. Graver, W. M. Carriere, E. E. Balkovich, and R. Thibodeau. Cost reporting elements and activity cost tradeoffs for defense system software (study results). Technical Report ESD-TR-77-262, Vol. 1, General Research Corporation, May 1977. 117
541. D. Gray, D. Bowes, N. Davey, Y. Sun, and B. Christianson. The misuse of the NASA metrics data program data sets for automated software defect prediction. In *15th Annual Conference on Evaluation & Assessment in Software Engineering 2011*, EASE 2011, pages 96–103, Apr. 2011. 364
542. J. Gray, C. Nyberg, M. Shah, and N. Govindaraju. Sort benchmark. <http://sortbenchmark.org>, July 2014. 337

543. W. D. Gray, C. R. Sims, W.-T. Fu, and M. J. Schoelles. The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, 113(3):461–482, 2006. [42](#)
544. M. Grechanik, C. McMillan, L. DeFerrari, M. Comi, S. Crespi, D. Poshyvanyk, C. Fu, Q. Xie, and C. Ghezzi. An empirical investigation into a large-scale Java open source code repository. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM ’10, pages 11:1–11:10, Sept. 2010. [165](#)
545. J. H. Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg, editor, *Universals of Language*, chapter 5, pages 58–90. MIT Press, 1963. [37](#)
546. H. I. Greenfield. An economist looks at data processing. *Computers and Automation*, 6(10):18–23, Oct. 1957. [77](#)
547. S. Greenstein. Did computer technology diffuse quickly?: Best and average practice in mainframe computers, 1968–1983. Working Paper No. 4647, National Bureau of Economic Research, Feb. 1994. [67](#)
548. Linux kernel statistics. website, June 2016. <https://www.github.com/gregkh/kernel-history>. [243](#), [269](#), [270](#)
549. C. Gregg and K. Hazelwood. Where is the data? Why you cannot debate CPU vs. GPU performance without the answer. In *IEEE International Symposium on Performance Analysis of Systems and Software*, ISPASS, pages 134–144, Apr. 2011. [321](#)
550. P. Grice. *Studies in the Way of Words*. Harvard University Press, 1989. [149](#)
551. D. A. Grier. The ENIAC, the verb "to program" and the emergence of digital computers. *IEEE Annals of the History of Computing*, 18(I):51–55, 1996. [1](#)
552. D. A. Grier. *When Computers were Human*. Princeton University Press, 2005. [1](#), [67](#)
553. S. Grimstad and M. Jørgensen. Inconsistency of expert judgement-based estimates of software development effort. *Journal of Systems and Software*, 80(11):1770–1777, Nov. 2007. [105](#)
554. E. Grochowski and R. E. Fontana, Jr. Future technology challenges for NAND flash and HDD products. Flash Memory Summit 2012, Santa Clara, CA, July 2012. [275](#)
555. U. Grömping. Relative importance for linear regression in r: The package relaimpo. *Journal of Statistical Software*, 17(1):1–27, Sept. 2006. [265](#)
556. E. H. B. M. Gronenschild, P. Habets, H. I. L. Jacobs, R. Mengelers, N. Rozendaal, J. van Os, and M. Marcelis. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS ONE*, 7(6):e38234, June 2012. [125](#)
557. H. R. J. Grosch. High speed arithmetic: The digital computer as a research tool. *Journal of the Optical Society of America*, 43(4):306–310, Apr. 1953. [1](#)
558. A. Grübler and N. Nakićenović. Long waves, technology diffusion, and substitution. Technical Report RP-91-17, International Institute for Applied Systems Analysis Laxenburg, Austria, Oct. 1991. [2](#)
559. W. Gruhl. Lessons learned cost/schedule assessment guide. Slides of talk, July 199x. [180](#)
560. M. Gubler. *Protean and boundaryless career orientations - an empirical study of IT professionals in Europe*. PhD thesis, Loughborough University, July 2011. [82](#)
561. L. Guerrouj, M. Di Penta, Y.-G. Guéhéneuc, and G. Antoniol. An experimental investigation on the effects of context on source code identifiers splitting and expansion. *Empirical Software Engineering*, 19(6):1706–1753, Dec. 2014. [312](#), [322](#)
562. H. S. Gunawi, M. Hao, T. Leesatapornwongsa, T. Patana-anake, T. Do, J. Adityatama, K. J. Eliazar, A. Laksono, J. F. Lukman, V. Martin, and A. D. Satria. What bugs live in the cloud? A study of 3000+ issues in cloud systems. In *Proceedings of the 5th ACM Symposium on Cloud Computing*, SOCC’14, pages 1–14, Nov. 2014. [335](#)
563. H. S. Gunawi, C. Rubio-González, A. C. Arpacı-Dusseau, R. H. Arpacı-Dusseau, and B. L. Liblit. EIO: Error handling is Occasionally correct. In *Proceedings of the 6<sup>th</sup> USENIX Conference on File and Storage Technologies*, pages 207–222, Feb. 2008. [139](#)
564. N. J. Gunther. A simple capacity model of massively parallel transaction systems. In *Proceedings of 19th International CMG Conference*, pages 1035–1044, Dec. 1993. [192](#)
565. N. J. Gunther. *Analysing Computer System Performance with Perl::PDQ*. Springer-Verlag, 2005. [192](#), [337](#)
566. J. Guo, K. Czarnecki, S. Apel, N. Siegmund, and A. Wąsowski. Variability-aware performance prediction: A statistical learning approach. In *IEEE/ACM 28th International Conference on Automated Software Engineering*, ASE 13, pages 301–311, Nov. 2013. [324](#)
567. O. Gurevich, M. A. Johnson, and A. E. Goldberg. Incidental verbatim memory for language. *Language and Cognition*, 2(1):45–78, May 2010. [155](#)
568. R. K. Guy. The strong law of small numbers. *American Mathematical Monthly*, 95(8):697–712, Oct. 1988. [221](#)
569. D. Hackenberg, R. Schöne, T. Ilsche, D. Molka, J. Schuchart, and R. Geyer. An energy efficiency feature survey of the Intel Haswell processor. In *International Parallel and Distributed Processing Symposium Workshop*, IPDPSW 2015, pages 896–904, May 2015. [338](#)
570. Hackerone. The 2018 hacker report. Technical report, hackerone, Dec. 2017. [83](#)
571. J. Haidt. *The Righteous Mind*. Vintage books, 2012. [35](#)
572. A. Halin, A. Nuttinck, M. Acher, X. Devroey, G. Perrouin, and B. Baudry. Test them all, is it worth it? A ground truth comparison of configuration sampling strategies. In *eprint arXiv:cs.SE/1710.07980*, Oct. 2017. [120](#), [143](#)
573. T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell. A systematic literature review on fault prediction performance in software engineering. *IEEE Transactions on Software Engineering*, 38(6):1276–1304, Nov. 2012. [364](#)
574. M. H. Halstead. A theoretical relationship between mental work and machine language programming. Technical Report CSD TR 67, Purdue University, Feb. 1972. [164](#)
575. M. H. Halstead and P. M. Zislis. Experimental verification of two theorems of software physics. Technical Report TR 97, Purdue University, June 1973. [164](#)
576. D. Z. Hambrick, F. L. Oswald, E. M. Altmann, E. J. Meinz, F. Gobet, and G. Campitelli. Deliberate practice: Is that all it takes to become an expert? *Intelligence*, 45(1):34–45, July-Aug. 2014. [31](#)
577. M. Hamill and K. Goseva-Popstojanova. Exploring the missing link: an empirical study of software fixes. *Software Testing, Verification and Reliability*, 24(8):684–705, Dec. 2014. [186](#), [187](#)
578. R. K. S. Hankin. Partial rank data with the hyper2 package: likelihood functions for generalized Bradley-Terry models. *The R Journal*, 9(2):429–439, 2017. [314](#)
579. J. E. Hannay, D. I. K. Sjøberg, and T. Dybå. A systematic review of theory use in software engineering experiments. *IEEE Transactions on Software Engineering*, 33(2):87–107, Feb. 2007. [5](#)
580. B. R. Harmon and N. I. Om. Schedule assessment methods for ballistic missile defense ground-based software development. IDA Paper P-3600, Institute for Defense Analysis, Aug. 2003. [107](#)
581. J. Haskel and S. Westlake. *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton University Press, 2018. [4](#), [53](#)
582. L. Hatton. *Safer C : Developing Software for High-integrity and Safety-critical Systems*. McGraw-Hill, 1995. [152](#)
583. L. Hatton. Reexamining the fault density-component size connection. *IEEE Software*, 14(2):89–97, Mar. 1997. [194](#)
584. L. Hatton. How accurately do engineers predict software maintenance tasks? *Computer*, 40(2):64–69, Feb. 2007. [182](#), [308](#)
585. R. Hatzinger and R. Dittrich. prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, 48(10):1–31, May 2012. [314](#)
586. M. D. Hauser, S. Carey, and L. B. Hauser. Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proceedings of the Royal Society B*, 267(1445):829–833, Apr. 2000. [16](#)
587. J. P. Haverty and R. L. Patrick. Programming languages and standardization in command and control. Research Memorandum RM-3447-PR, The RAND Corporation, Jan. 1963. [89](#)
588. D. M. Hawkins. *Identification of Outliers*. Springer, 1980. [365](#)
589. G. Hawkins, S. D. Brown, M. Steyvers, and E.-J. Wagenmakers. Context effects in multi-alternative decision making: Empirical data and a bayesian model. *Cognitive Science*, 36(3):498–516, Apr. 2012. [49](#)

590. G. E. Hawkins, S. D. Brown, M. Steyvers, and E.-J. Wagenmakers. An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic Bulletin & Review*, 19(2):339–348, Apr. 2012. [322](#)
591. J. A. Hawkins. *Efficiency and Complexity in Grammars*. Oxford University Press, 2007. [155](#)
592. B. Hayes. Third base. *American Scientist*, 89(6):490–494, 2001. [4](#)
593. S. Hazelhurst. Truth in advertising: Reporting performance of computer programs, algorithms and the impact of architecture and systems environment. *South African Computer Journal*, 46:24–37, Dec. 2010. [274](#)
594. S. Head and J. Nelson. Data rights valuation in software acquisitions. Technical Report DRM-2012-001825-Final, CNA Analysis & Solutions, Sept. 2012. [92](#)
595. A. Heathcote, S. Brown, and D. J. K. Mewhort. The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7(2):185–207, Apr. 2000. [24](#)
596. R. Heeks. The uneven profile of Indian software exports. Working Paper No. 3, University of Manchester, Oct. 1998. [70](#)
597. D. H. Helmer, S. Mackay, K. Selvey-Clinton, R. Yoon, and H. Furukawa. Worldwide capital and fixed assets guide 2016. Technical Report EYG no. DL1528, EYGM Limited, 2016. [60](#)
598. D. R. Helsel. *Statistics for Censored Environmental Data using Minitab and R*. John Wiley & Sons, second edition, 2012. [295](#)
599. A. Hemel and R. Koschke. Reverse engineering variability in source code using clone detection-A case study for Linux variants of consumer electronic devices. In *19th Working Conference on Reverse Engineering*, WCRE’12, pages 357–366, Oct. 2012. [84](#)
600. A. Henik and J. Tzelgov. Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, 10(4):389–395, 1982. [41](#)
601. T. Henkel. Hardware roundup. *ComputerWorld*, XIX(33):23–34, Aug. 1985. [78](#)
602. J. Henrich. How adaptive cultural processes can produce maladaptive losses—The Tasmanian case. *American Antiquity*, 69(2):197–214, Apr. 2004. [71](#)
603. J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? Working Paper No. 139, German Data Forum (RatSWD), Apr. 2010. [17](#)
604. J. Henrich and R. McElreath. Are peasants risk-averse decision makers? *Current Anthropology*, 43(1):172–181, Feb. 2002. [44](#)
605. F. Hermans and E. Murphy-Hill. Enron’s spreadsheets and related emails: A dataset and analysis. In *Proceedings of the 37th International Conference on Software Engineering-Volume 2*, pages 7–16, May 2015. [150](#)
606. T. Herr, B. Schneier, and C. Morris. Taking stock: Estimating vulnerability rediscovery. Paper, Belfer Center for Science and International Affairs, Harvard Kennedy School, July 2017. [134](#)
607. I. Herrera, G. Robles, J. M. González-Barahona, A. Capiluppi, and J. F. Ramil. Comparison between SLOCs and number of files as size metrics for software evolution analysis. In *Proceedings of the 10th European Conference on Software Maintenance and Reengineering*, CSMR 2006, pages 206–213, Mar. 2006. [151](#)
608. E. Herrmann, J. Call, M. Victoria, Hernández-Lloreda, B. Hare, and M. Tomasello. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, 317(5843):1360–1366, Sept. 2007. [43](#), [71](#)
609. R. Hersh. *18 Unconventional Essays on the Nature of Mathematics*. Springer, 2006. [124](#)
610. K. Herzig, S. Just, and A. Zeller. It’s not a bug, it’s a feature: How misclassification impacts bug prediction. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE’13, pages 392–401, May 2013. [125](#), [365](#)
611. K. Herzig and A. Zeller. Untangling changes. Submitted to MSR 2013, 2013. [365](#)
612. T. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. In *eprint arXiv:stat.OT/1411.5279*, Nov. 2014. [231](#)
613. M. Hicks, C. O’Malley, S. Nichols, and B. Anderson. Comparison of 2D and 3D representations for visualising telecommunication usage. *Behaviour & Information technology*, 22(3):185–201, May 2003. [188](#)
614. E. T. Higgins. Value from regulatory fit. *Current Directions in Psychological Science*, 14(4):209–213, 2005. [19](#)
615. N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 1996. [123](#)
616. M. Hilbert and P. López. Supporting online material for: The world’s technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, Apr. 2011. [78](#)
617. J. Hill, L. Thomas, and D. E. Allen. Experts’ estimates of task durations in software development projects. *International Journal of Project Management*, 18(1):13–21, Feb. 2000. [108](#)
618. T. T. Hills, P. M. Todd, and M. N. Jones. Foraging in semantic fields: How we search through memory. *Topics in Cognitive Science*, 7(3):513–534, July 2015. [24](#)
619. D. J. Hilton. The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, 118(2):248–271, 1995. [36](#)
620. A. Hindle, M. W. Godfrey, and R. C. Holt. Reading beside the lines: Indentation as a proxy for complexity metrics. In *The 16th IEEE International Conference on Program Comprehension*, ICPC 2008, pages 133–142, June 2008. [287](#), [288](#)
621. T. Hirao, A. Ihara, Y. Ueda, P. Phannachitta, and K. ichi Matsumoto. The impact of a low level of agreement among reviewers in a code review process. In *IFIP International Conference on Open Source Systems*, OSS 2016, pages 97–110, May-June 2016. [145](#)
622. S. C. Hirtle and J. Jonides. Evidence for hierarchies in cognitive maps. *Memory & Cognition*, 13(3):208–217, 1985. [41](#)
623. C. A. R. Hoare. An axiomatic basis for computer programming. *Communications of the ACM*, 12(10):576–583, Oct. 1969. [124](#)
624. M. Hocko and T. Kalibera. Reducing performance non-determinism via cache-aware page allocation strategies. In *Proceedings of the First Joint WOSP/SIPEW International Conference on Performance Engineering*, WOSP/SIPEW’10, pages 223–234, Jan. 2010. [342](#)
625. A. Höfer. Exploratory comparison of expert and novice pair programmers. *Computing and Informatics*, 29(1):73–91, 2010. [261](#)
626. D. D. Hoffman. *Visual Intelligence: How We Create What We See*. W. W. Norton, 2000. [15](#), [32](#)
627. R. Hofman. *Behavioral Products Quality Assessment Model on the Software Market*. PhD thesis, Poznan University of Economics, Oct. 2011. [47](#)
628. J. Hofmeister, J. Siegmund, and D. V. Holt. Shorter identifier names take longer to comprehend. In *24th International Conference on Software Analysis, Evolution and Reengineering*, SANER 2017, pages 217–227, Feb. 2017. [163](#)
629. G. Hofstede. *Culture’s Consequences: International Differences in Work-Related Values*. Sage Publications, abridged edition, 1984. [71](#)
630. R. M. Hogarth and H. J. Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–55, Jan. 1992. [26](#), [27](#)
631. R. M. Hogarth, C. R. M. McKenzie, B. J. Gibbs, and M. A. Marquis. Learning from feedback: Exactness and incentives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(4):734–752, 1991. [31](#), [75](#)
632. B. D. Holbrook and W. S. Brown. A history of computing research at Bell Laboratories (1937–1975). Computing Science Technical Report No. 99, AT&T Bell Laboratories, 1982. [67](#)
633. M. Holdway. An alternative methodology: Valuing quality change for microprocessors in the PPI. In *Issues in Measuring Price Change and Consumption*. Bureau of Labor Statistics, June 2000. [4](#)
634. W. B. Holland. Soviet cybernetics technology: viii. Report on the algorithmic language ALGEC (final version). Research Memorandum RM-5136-PR, The RAND Corporation, Dec. 1966. [73](#)
635. J. K. Hollmann. Estimate accuracy: Dealing with reality. *Cost Engineering Journal*, 54(6):17–27, Nov.-Dec. 2012. [105](#)
636. A. A. Hook, B. Brykczynski, C. W. McDonald, S. H. Nash, and C. Youngblut. A survey of computer programming languages currently used in the Department of Defense. IDA PAPER P-3054, Institute for Defense Analyses, Jan. 1995. [90](#)
637. R. Hoosain. Correlation between pronunciation speed and digit span size. *Perception and Motor Skills*, 55:1128–1128, 1982. [322](#)

638. R. Hoosain and F. Salili. Language differences, working memory, and mathematical ability. In M. M. Grunberg, P. E. Morris, and R. N. Sykes, editors, *Practical aspects of memory: Current research and issues*, volume 2, pages 512–517. John Wiley & Sons, Inc, 1988. 21
639. M. Hoppe and S. Hanenberg. Do developers benefit from generic types? An empirical comparison of generic and raw types in Java. In *Proceedings of the 2013 ACM SIGPLAN international conference on Object oriented programming systems languages & applications*, OOPSLA’13, pages 457–474, Oct. 2013. 154
640. W. Hordijk, M. L. Poncino, and R. Wieringa. Harmfulness of code duplication a structured review of the evidence. In *13<sup>th</sup> International Conference on Evaluation and Assessment in Software Engineering*, EASE’09, pages 88–97, Apr. 2009. 96
641. M. R. Horton. *Portable C Software*. Prentice-Hall, Inc, Upper Saddle River, NJ 07458, USA, 1990. 152
642. A. D. Householder and J. M. Foote. Probability-based parameter selection for black-box fuzz testing. Technical Note CMU/SEI-2012-TN-019, Software Engineering Institute, Carnegie Mellon University, Aug. 2012. 132
643. M. W. Howard and M. J. Kahana. Context variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 25(4):923–941, 1999. 24
644. J. Howison and J. B. Herbsleb. Incentives and integration in scientific software production. In *Proceedings of the 2013 conference on Computer supported cooperative work*, CSCW’13, pages 459–470, Mar. 2013. 60, 69, 81
645. L. Hribar, S. Bogovac, and Z. Marinčić. Implementation of fault slip through in design phase of the project. In *miproBIS 2008: International Conference on Business Intelligence Systems*, page ???, May 2008. 144
646. X. Huang, J. Xie, N. O. Otecko, and M. Peng. Accessibility and update status of published software: Benefits and missed opportunities. *Frontiers in Research Metrics and Analytics*, 2(???):???, Feb. 2017. 120
647. B. A. Huberman. The dynamics of organizational learning. *Computational & Mathematical Organization Theory*, 7(2):145–153, Aug. 2001. 76
648. H. Huijgens and R. van Solingen. Measuring best-in-class software releases. In *Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, pages 137–146, Oct. 2013. 107
649. H. Huijgens and F. Vogezelang. Do estimators learn? On the effect of a positively skewed distribution of effort data on software portfolio productivity. Technical Report TUD-SERG-2016-004, Delft University of Technology, 2016. 58
650. C. Hulme, S. Maughan, and G. D. A. Brown. Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, 30(6):685–701, 1991. 22
651. C. R. Hulten. Decoding Microsoft: Intangible capital as a source of company growth. Working Paper 15799, National Bureau of Economic Research, USA, Mar. 2010. 58
652. R. Hundt, E. Raman, M. Thuresson, and N. Vachharajani. MAO—an extensible micro-architectural optimizer. In *Proceedings of the 9th Annual IEEE/ACM International Symposium on Code Generation and Optimization*, CGO’11, pages 1–10, Apr. 2011. 339
653. S. Hunold and A. Carpen-Amarie. MPI benchmarking revisited: Experimental design and reproducibility. In *eprint arXiv:cs.DC/1505.07734v3*, Sept. 2015. 338
654. S. Hunold, A. Carpen-Amarie, and J. L. Träff. Reproducible MPI micro-benchmarking isn’t as easy as you think. In *Proceedings of the 21st European MPI Users’ Group Meeting*, EuroMPI/ASIA’14, pages 69–76, Sept. 2014. 209, 210
655. E. Hunt. The Whorfian hypothesis: A cognitive psychology perspective. *Psychological Review*, 98(3):377–389, 1991. 166
656. M. J. Hurlstone, G. J. Hitch, and A. D. Baddeley. Memory for serial order across domains: An overview of the literature and directions for future research. *Psychonomic Bulletin & Review*, 140(2):229–373, Mar. 2014. 24
657. A. A. Hwang, I. A. Stefanovici, and B. Schroeder. Cosmic rays don’t strike twice: Understanding the nature of DRAM errors and the implications for system design. In *Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS XVII, pages 111–122, Mar. 2012. 142
658. IBM. Specifications for the IBM mathematical FORMula TRANSlating system, FORTRAN. Programming Research Group, Applied Science Division, International Business Machines Corporation, Nov. 1954. 89
659. R. Ierusalimschy, L. H. de Figueiredo, and W. Celes. The evolution of Lua. In *Proceedings of the third ACM SIGPLAN conference on History of programming languages*, HOPL III, pages 1–26, June 2007. 90
660. J. Iivonen. Identifying and characterizing highly performing testers—A case study in three software product companies. Thesis (m.s.), Helsinki University of Technology, Department of Computer Science and Engineering, Oct. 2009. 51
661. I. Imbo and J.-A. LeFevre. Cultural differences in complex addition: Efficient Chinese versus adaptive Belgians and Canadians. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6):1465–1476, Nov. 2009. 40
662. I. Imbo, A. Vandierendonck, and E. Vergauwe. The role of working memory in carrying and borrowing. *Psychological Research*, 71(4):467–483, July 2007. 40
663. L. Inozemtseva and R. Holmes. Coverage is not strongly correlated with test suite effectiveness. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE’14, pages 435–445, June 2014. 147
664. Intel. 6th generation Intel processor family. Specification update 332689-010EN, Intel Corporation, Apr. 2017. 136
665. J. P. A. Ioannidis. Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2):218–228, July 2005. 8
666. J. P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, Aug. 2005. 229
667. A. Iosup, M. Jan, O. Sonmez, and D. H. J. Epema. On the dynamic resource availability in grids. Rapport de recherche no 6172, INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE, Apr. 2007. 143
668. G. Irlam. Unix file size survey-1993. <http://www.base.com/gordoni/ufs93.html>, Sept. 1993. 211
669. F. Irving. Github users since service started. [https://classic.scrapewiki.com/scrapers/github\\_users\\_each\\_year/](https://classic.scrapewiki.com/scrapers/github_users_each_year/), Mar. 2016. 63
670. ISO. *ISO/IEC Guide 25:1990 General requirements for the competence of calibration and testing laboratories*. International Organization for Standardization, 1990. 146
671. ISO SC22. *ISO/IEC 18009:1999 Information technology – Programming languages – Ada: Conformity assessment of a language processor*. International Organization for Standardization, 1990. Last reviewed and confirmed in 2015. 146
672. ISO SC22. *ISO/IEC 13210:1999 Information technology – Requirements and guidelines for test methods specifications and test method implementation for measuring conformance to POSIX standards*. International Organization for Standardization, 1999. 146
673. A. Israeli and D. G. Feitelson. The Linux kernel as a case study in software evolution. *Journal of Systems and Software*, 83(3):485–501, Mar. 2010. 164, 246, 247, 252, 275
674. International telecommunication union. website, July 2012. <http://www.itu.int>. 136
675. R. K. Iyer, S. E. Butner, and E. J. McCluskey. An exponential failure/load relationship: Results of a multi-computer statistical study. Technical Report CRC-81-6, Computer Systems Laboratory, Stanford University, Aug. 1981. 126
676. M. Y. Jaber. Learning and forgetting models and their applications. In A. B. Badiru, editor, *Handbook of Industrial and Systems Engineering*, chapter 30. CRC Press-Taylor & Francis Group, Dec. 2005. 26
677. A. N. Jackson. Formats over time: Exploring UK web history. In *eprint arXiv:cs.DL/1210.1714v1*, Oct. 2012. 87
678. J. Jacobs and B. Rudis. *Data-Driven Security*. John Wiley & Sons, Inc, 2014. 372
679. R. Jaeschke. *Portability and the C Language*. Hayden Books, 4300 West 62nd Street, Indianapolis, IN 46268, USA, 1989. 152

680. L. R. Jager and J. T. Leek. Empirical estimates suggest most published research is true. *Biostatistics*, 15(1):1–12, 2014. 229
681. B. Jamtveit, E. Jettestuen, and J. Mathiesen. Scaling properties of European research units. *PNAS*, 106(32):13160–13163, Aug. 2009. 82
682. A. R. Jansen. *Encoding and Parsing of Algebraic Expressions by Experienced Users of Mathematics*. PhD thesis, School of Computer Science and Software Engineering, Monash University, Jan. 2002. 34
683. A. R. Jansen, K. Marriott, and G. W. Yelland. Parsing of algebraic expressions by experienced users of mathematics. *European Journal of Cognitive Psychology*, 19(2):286–320, 2007. 34
684. C. J. M. Jansen and M. M. W. Pollmann. On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics*, 8(3):187–201, 2001. 137
685. J. J. Jenkins. Remember that old theory of memory? Well, forget it! *American Psychologist*, 29(11):785–795, 1974. 155
686. Computer smartphone and tablet marketshare: 1975–2012. website, Dec. 2012. <http://jeremyreimer.com/m-item.lsp?i=137>. 3, 68
687. J. Jiang, D. Lo, J. He, X. Xia, P. S. Kochhar, and L. Zhang. Why and how developers fork what from whom in GitHub. *Empirical Software Engineering*, 22(1):547–578, Feb. 2017. 86
688. Introducing open salaries at buffer our transparent formula and all individual salaries buffer. website, Dec. 2013. <https://open.buffer.com/introducing-open-salaries-at-buffer-including-our-transparent-formula-and-all-individual-salaries/>. 83
689. The life cycle of a cpu. website: accessed 13 Jul 2017, 2010. <http://www.cpushack.com/life-cycle-of-cpu.html>. 221
690. Clark’s sector model for US economy 1850–2009. website, 2011. <http://www.63alfred.com/whomakesit/clarksmodel.htm>. 81
691. D. D. P. Johnson, N. B. Weidmann, and L.-E. Cederman. Fortune favours the bold: An agent-based model reveals adaptive advantages of overconfidence in war. *PLoS ONE*, 6(6):e20851, Apr. 2011. 48
692. L. Johnson. Applied data research inc. (ADR). Technical report, Computer History Museum, Feb. 2010. 79
693. P. M. Johnson and A. M. Disney. A critical analysis of PSP data quality: Results from a case study. *Empirical Software Engineering*, 4(4):317–349, Dec. 1999. 364
694. W. L. Johnson. *Intention-Based Diagnosis of Novice Programming Errors*. Morgan Kaufmann Publishers, Inc, 1986. 165
695. C. I. Jones. The facts of economic growth. In J. B. Taylor and H. Uhlig, editors, *Handbook of Macroeconomics, Volume 2A*, chapter 1, pages 3–69. Elsevier B. V., Nov. 2016. 4
696. D. Jones. Why userspace sucks—or 101 really dumb things your app shouldn’t do. In *Proceedings of the Linux Symposium: Volume One*, pages 441–450, July 2006. 342
697. D. M. Jones. Who guards the guardians? [www.knosof.co.uk/whoguard.html](http://www.knosof.co.uk/whoguard.html), 1992. 148
698. D. M. Jones. The  $7 \pm 2$  urban legend. MISRA C 2002 conference <http://www.knosof.co.uk/cbook/misart.pdf>, Oct. 2002. 20, 322
699. D. M. Jones. The new C Standard: An economic and cultural commentary. Knowledge Software, Ltd, 2005. 74, 78, 91, 117, 139, 160, 162, 165, 169, 170, 172, 174, 175, 190, 215, 258, 322, 371
700. D. M. Jones. Developer beliefs about binary operator precedence. *C Vu*, 18(4):14–21, Aug. 2006. 20, 44, 263, 314, 322, 346
701. D. M. Jones. Operand names influence operator precedence decisions. *C Vu*, 20(1):5–11, Feb. 2008. 44, 160, 346
702. D. M. Jones. Deciding between if and switch when writing code. *C Vu*, 21(5):14–20, Nov. 2009. 169
703. D. M. Jones. Developer characterization of data structure fields decisions. *C Vu*, 20(6):14–18, Jan. 2009. 30, 215, 216, 312, 313, 346
704. D. M. Jones. Effects of risk attitude on recall of assignment statements. *C Vu*, 23(6):19–22, Jan. 2012. 44
705. D. M. Jones. Code & data for Evidence-based software engineering using R. <http://www.github.com/Derek-Jones/ESEUR>, 2018. 2
706. M. N. Jones and D. J. K. Mewhort. Case-sensitive letter and bigram frequency counts from large-scale English corpora. *Behavior Research Methods, Instruments, & Computers*, 36(3):388–396, 2004. 164
707. R. Jongeling, S. Datta, and A. Serebrenik. Choosing your weapons: On sentiment analysis tools for software engineering research. In *31st International Conference on Software Maintenance and Evolution, ICSME 2015*, pages 531–535, Sept. 2015. 316
708. M. R. Jongerden. *Model-based energy analysis of battery powered systems*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, Dec. 2010. 339
709. M. Jørgensen. An empirical study of software maintenance tasks. *Software Maintenance: Research and Practice*, 7(1):27–48, Jan. 1995. 261
710. M. Jørgensen. Regression models of software development effort estimation accuracy and bias. *Empirical Software Engineering*, 9(4):297–394, Apr. 2004. 105, 108
711. M. Jørgensen. Better selection of software providers through trial-sourcing. *IEEE Software*, 33(5):48–53, Sept.–Oct. 2016. 108, 109
712. M. Jørgensen. A survey on the characteristics of projects with success in delivering client benefits. *Information and Software Technology*, 78:83–94, Oct. 2016. 117
713. M. Jørgensen and G. J. Carelius. An empirical study of software project bidding. *IEEE Transactions on Software Engineering*, 30(12):953–969, Dec. 2004. 110, 330
714. M. Jørgensen, T. Dybå, K. Liestøl, and D. I. K. Sjøberg. Incorrect results in software engineering experiments: How to improve research practices. *Journal of Systems and Software*, 116:133–145, June 2016. 229
715. M. Jørgensen and S. Grimstad. Software development estimation biases: The role of interdependence. *IEEE Transactions on Software Engineering*, 38(3):677–693, May 2012. 17, 51
716. M. Jørgensen, U. Indahl, and D. I. K. Sjøberg. Software effort estimation by analogy and “regression toward the mean”. *Journal of Systems and Software*, 68(3):253–262, Dec. 2003. 245
717. M. Jørgensen and K. Moløkken. Eliminating over-confidence in software development effort estimates. In F. Bomarius and H. Iida, editors, *Product Focused Software Process Improvement*, volume 3009 of *Lecture Notes in Computer Science*, pages 174–184. Springer Berlin Heidelberg, Apr. 2004. 332
718. M. Jørgensen and K. Moløkken. Understanding reasons for errors in software effort estimates. *IEEE Transactions on Software Engineering*, 30(12):993–1007, Dec. 2004. 42
719. M. Jørgensen and K. Moløkken. How large are software cost overruns? A review of the 1994 CHAOS report. *Journal Information and Software Technology*, 48(4):297–301, Apr. 2006. 104
720. M. Jørgensen and D. I. K. Sjøberg. The impact of customer expectation on software development effort estimates. *International Journal of Project Management*, 22(4):317–325, May 2004. 51, 105
721. M. Jørgensen and D. I. K. Sjøberg. Learning from experience in a software maintenance environment. *Journal of Computer Science*, 1(4):538–542, Apr. 2005. 253
722. D. Joseph, W. F. Boh, S. Ang, and S. A. Slaughter. The career paths less (or more) travelled: A sequence analysis of IT career histories, mobility patterns, and career success. *MIS Quarterly*, 36(2):427–452, June 2012. 82
723. S. Kahrs. Mistakes and ambiguities in the definition of standard ML. LFCS report ECS-LFCS-93-257, University of Edinburgh, Scotland, Apr. 1993. 124
724. J. W. Kalat. *Biological Psychology*. Wadsworth, seventh edition, 2001. 16
725. T. Kalibera, L. Bulej, and P. Tůma. Benchmark precision and random initial state. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2005*, pages 853–862. Society for Modeling and Simulation (SCS), July 2005. 342
726. A. Kaltenbrunner, V. Gómez, A. Moghnieh, R. Meza, J. Blat, and V. López. Homogeneous temporal activity patterns in a large online communication space. In *eprint arXiv:cs.NI/0708.1579*, Aug. 2007. 210
727. D. Kaminsky, M. Eddington, and A. Cecchetti. Showing how security has (and hasn’t) improved, after ten years of trying. CanSecWest Applied Security Conference, Dec. 2011. 133, 134
728. T. Kamiya. How code skips over revisions. In *Proceedings of the 5th International Workshop on Software Clones, IWSC 2011*, pages 69–70, May 2011. 166

729. V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 49(11-12):1073–1086, Apr. 2007. 5
730. P. Kampstra and C. Verhoef. Reliability of function point counts. ???, ???(??):???, Apr. 200? 107
731. P. Kampstra and C. Verhoef. Benchmarking the expected loss of a federal IT portfolio. ???, July 2009. 241
732. C. Kaner. Liability for defective documentation. In *Proceedings of the 21st annual international conference on Documentation*, SIGDOC'03, pages 192–197, Oct. 2003. 88, 141
733. C. Kaner and D. Pels. *Bad Software: What To Do When Software Fails*. John Wiley & Sons, Inc, 1998. 128
734. Y. Kang, B. Ray, and S. Jana. APEx: Automated inference of error specifications for C APIs. In *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering*, ASE 2016, pages 472–482, Sept. 2016. 147
735. D. Karlis and E. Xekalaki. Mixed poisson distributions. *International Statistical Review*, 73(1):35–58, Apr. 2005. 203
736. J. Karlsson and K. Ryan. A cost-value approach for prioritizing requirements. *IEEE Software*, 14(5):67–74, Sept. 1997. 118
737. D. Kawrykow and M. P. Robillard. Non-essential changes in version histories. In *Proceedings of the 33rd International Conference on Software Engineering*, ICSE'11, pages 351–360, May 2011. 95
738. M. Kazandjieva, B. Heller, O. Gnawali, P. Levis, and C. Kozyrakis. Green enterprise computing data: Assumptions and realities. In *Proceedings of the 2012 International Green Computing Conference*, IGCC'12, pages 1–10, June 2012. 78
739. M. Keil and D. Robey. Blowing the whistle on troubled software projects. *Communications of the ACM*, 44(4):87–93, Apr. 2001. 111
740. P. Keil, J. M. Bennett, B. Bourgeois, G. E. G.-P. na, A. A. M. MacDonald, C. Meyer, K. S. Ramirez, and B. Yguel. From computer operating systems to biodiversity: co-emergence of ecological and evolutionary patterns. *PNAS*, ???(??):???, Aug. 2016. 69
741. C. F. Kemerer. An empirical validation of software cost estimation models. *Communications of the ACM*, 30(5):416–429, May 1987. 106
742. Z. Kenessey. The primary, secondary, tertiary and quaternary sectors of the economy. *The Review of Income and Wealth*, 33(4):359–385, Dec. 1987. 81
743. D. O. Kennedy and A. B. Scholey. Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology*, 149(1):63–71, May 2000. 50
744. B. W. Kernighan and R. Pike. *The Practice of Programming*. Addison-Wesley, 1999. 152
745. M. D. Kernighan, K. W. Church, and W. A. Gale. A spelling correction program based on a noisy channel model. In *Proceedings of COLING-90*, pages 205–210, 1990. 162
746. E. Keuleers, P. Lacey, K. Rastle, and M. Brysbaert. The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior and Research Methods*, 44(1):287–304, Mar. 2012. 24
747. H. Khalid, M. Nagappan, E. Shihab, and A. E. Hassan. Prioritizing the devices to test your app on: A case study of Android game apps. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2014, pages 610–620, Nov. 2014. 60
748. L. M. Khan. Amazon’s antitrust paradox. *The Yale Law Journal*, 126(3):710–805, Jan. 2017. 98
749. P.-V. Khuong and P. Morin. Array layouts for comparison-based searching. In *eprint arXiv:cs.DS/1509.05053*, Mar. 2017. 341
750. P. D. Killworth and H. R. Bernard. Informant accuracy in social network data. *Human Organization*, 35(3):269–286, 1976. 318
751. D. Kim, E. Murphy-Hill, C. Parnin, C. Bird, and R. Garciaad. The reaction of open-source projects to new language features: An empirical study of C# generics. *The Journal of Object Technology*, 12(4):1–26, Nov. 2013. 167
752. H. Kim. *Informed Storage Management for Mobile Platforms*. PhD thesis, College of Computing, Georgia Institute of Technology, Dec. 2012. 341
753. S. Kim. The classification of information and communication technology investment in financial accounting. Thesis (m.s.), School of Information Technologies, University of Sydney, 2013. 60
754. K. Kina, M. Tsunoda, H. Hata, H. Tamada, and H. Igaki. Analyzing the decision criteria of software developers based on prospect theory. In *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering*, SANER, pages 644–648, Mar. 2016. 48
755. J. King and M. A. Just. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30:580–602, 1991. 22
756. W. Kintsch. *Comprehension: A paradigm for cognition*. Cambridge University Press, 1998. 158
757. W. Kintsch and J. Keenan. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5(3):257–274, Nov. 1973. 155
758. W. Kintsch, E. Kozminsky, W. J. Streby, G. McKoon, and J. M. Keenan. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14(2):196–214, Apr. 1975. 155
759. W. Kintsch, T. S. Mandel, and E. Kozminsky. Summarizing scrambled stories. *Memory & Cognition*, 5(5):547–552, 1977. 158
760. K. N. Kirby and R. J. Herrnstein. Preference reversals due to myopic discounting of delayed reward. *Psychological Science*, 6(2):83–89, Mar. 1995. 49
761. D. Kirsh and P. Maglio. On distinguishing epistemic from pragmatic action. *Cognitive Science*, 18(4):513–549, Oct. 1994. 18
762. L. B. Kish. Moore’s law and the energy requirement of computing versus performance. *IEE Proceedings-Circuits, Devices and Systems*, 151(2):190–194, Apr. 2004. 141
763. S. Kitayama and M. Karasawa. Implicit self-esteem in Japan: Nameletters and birthday numbers. *Personality & Social Psychology Bulletin*, 23(7):736–742, 1997. 73
764. B. Kitchenham, S. L. Pfleeger, B. McColl, and S. Eagan. An empirical study of maintenance and development estimation accuracy. *The Journal of Systems and Software*, 64(1):57–77, Oct. 2002. 105, 108
765. B. A. Kitchenham and N. R. Taylor. Software project development cost estimation. *The Journal of Systems and Software*, 5(4):267–278, Nov. 1985. 117
766. D. Klahr, W. G. Chase, and E. A. Lovelace. Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 9(3):462–477, 1983. 24
767. J. Klayman and Y.-W. Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211–228, 1987. 52
768. B. Klein. The decision making problem in development. In Universities-National Bureau Committee for Economic Research, Committee on Economic Growth of the Social Science Research Council, editor, *The Rate and Direction of Inventive Activity: Economic and Social Factors*, chapter 19, pages 477–508. Princeton University Press, 1962. 111
769. S. B. Klein, L. Cosmides, J. Tooby, and S. Chance. Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review*, 109(2):306–329, 2002. 19
770. S. Kleinschmager, R. Robbes, A. Stefk, S. Hanenberg, and É. Tanner. Do static type systems improve the maintainability of software systems? An empirical study. In *20th International Conference on Program Comprehension*, ICPC'12, pages 153–162, June 2012. 153
771. S. Klepper and K. L. Simons. Technological extinctions of industrial firms: An inquiry into their nature and causes. *Industrial and Corporate Change*, 6(2):379–460, Mar. 1997. 79
772. P. Klint, D. Landman, and J. Vinju. Exploring the limits of domain model recovery. In *29th IEEE International Conference on Software Maintenance*, ICSM'13, pages 120–129, Sept. 2013. 109
773. K. E. Knight. Changes in computer performance. *Datamation*, 12(9):40–54, Sept. 1966. 336
774. K. E. Knight. Evolving computer performance 1963–1967. *Datamation*, 14(1):31–35, Jan. 1968. 5, 336
775. D. E. Knuth. The errors of TeX. *Software—Practice and Experience*, 19(7):607–685, 1989. 126

776. D. Kobak, S. Shpilkin, and M. S. Pshenichnikov. Integer percentages as electoral falsification fingerprints. In *eprint arXiv:stat.AP/1410.6059v4*, June 2016. 137
777. A. Koenig. *C Traps and Pitfalls*. Addison-Wesley, 1989. 152
778. R. Kohavi and R. Longbotham. Unexpected results in online controlled experiments. *ACM SIGKDD Explorations Newsletter*, 12(2):31–35, Dec. 2010. 322
779. P. A. Kolers. Reading A year later. *Journal of Experimental Psychology: Human Learning and Memory*, 2(3):554–565, 1976. 160
780. P. A. Kolers and D. N. Perkins. Spatial and ordinal components of form perception and literacy. *Cognitive Psychology*, 7(2):228–267, Apr. 1975. 159
781. J. G. Koomey, S. Berard, M. Sanchez, and H. Wong. Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, 33(3):46–54, July-Sept. 2011. 3
782. A. Koriat. How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, 100(4):609–639, 1993. 23
783. A. G. Koru, K. El Emam, D. Zhang, H. Liu, and D. Mathew. Theory of relative defect proneness: Replicated studies on the functional form of the size-defect relationship. *Empirical Software Engineering*, 13(5):473–498, Oct. 2008. 140, 302
784. S. M. Kosslyn. *Graph Design for the Eye and Mind*. Oxford University Press, 2006. 192
785. S. M. Kosslyn and S. P. Shwartz. Empirical constraints on theories of visual imagery. In J. Long and A. D. Baddeley, editors, *Attention and Performance IX*, pages 241–260. Lawrence Erlbaum Associates, 1981. 18
786. E. Krevat, J. Tucek, and G. R. Ganger. Disks are like snowflakes: No two are alike. In *Proceedings of the 13th USENIX conference on Hot topics in operating systems*, HotOS’13, May 2013. 340
787. E. C. Kubie. Recollections of the first software company. *IEEE Annals of the History of Computing*, 16(2):65–71, June 1994. 79
788. M. Kubovy and M. van den Berg. The whole is equal to the sum of its parts: A probabilistic model of grouping by proximity and similarity in regular patterns. *Psychological Review*, 115(1):131–154, 2008. 32, 33
789. D. R. Kuhn and D. R. Wallace. Software fault interactions and implications for software testing. *IEEE Transactions on Software Engineering*, 30(6):418–421, June 2004. 146
790. M. Kuhrmann, C. Konopka, P. Nellemann, P. Diebold, and J. Münch. Software process improvement: Where is the evidence? In *Proceedings of the 2015 International Conference on Software and System Process*, pages 107–116, Aug. 2015. 6
791. R. G. Kula, D. M. German, A. Ouni, T. Ishio, and K. Inoue. Do developers update their library dependencies? An empirical study on the impact of security advisories on library migration. In *eprint arXiv:cs.SE/1709.04621*, Sept. 2017. 9, 92
792. R. Kumar. The business of scaling. *IEEE Solid-State Circuits Society Newsletter*, 12(1):22–26, 2007. 4
793. G. Kunda. *Engineering Culture: Control and Commitment in a High-Tech Corporation*. Temple University Press, 1992. 80
794. P. Küngas, S. Vakulenko, M. Dumas, C. Parra, and F. Casati. Reverse-engineering conference rankings: What does it take to make a reputable conference? *Scientometrics*, 96(2):651–665, Aug. 2013. 8
795. G. Kunst. Language popularity. <http://langpop.corger.nl/results>, 2013. 248
796. R. Kurzban, A. Duckworth, J. W. Kable, and J. Myers. An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, 36(6):661–679, Dec. 2013. 43
797. D. S. Kusumo, M. Staples, L. Zhu, and R. Jeffery. Analyzing differences in risk perceptions between developers and acquirers in OTS-based custom software projects using stakeholder analysis. In *Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM’12, pages 69–78, Sept. 2012. 112
798. T. Labiner. A big decision: Lease or buy? *Computers and Automation*, 6(10):6–8, Oct. 1957. 76, 77
799. W. Labov. The boundaries of words and their meaning. In C.-J. N. Bailey and R. W. Shuy, editors, *New ways of analyzing variation of English*, pages 340–373. Georgetown Press, 1973. 30
800. W. Labov. *Principles of Linguistic Change, volume 3: Cognitive and Cultural Factors*. Wiley-Blackwell, 2010. 160
801. J. C. Lagarias. *The Kepler Conjecture: The Hales-Ferguson Proof by Thomas C. Hales Samuel P. Ferguson*. Springer, 2010. 124
802. K. Laitinen. *Natural naming in software development and maintenance*. PhD thesis, University of Oulu, Finland, Oct. 1995. 161
803. G. Lakoff and M. Johnson. *Metaphors We Live By*. The University of Chicago Press, 1980. 38, 72
804. B. L. Lambert, K.-Y. Chang, and P. Gupta. Effects of frequency and similarity neighborhoods on pharmacists’ visual perception of drug names. *Social Science and Medicine*, 57(10):1939–1955, Nov. 2003. 162
805. R. Lämmel, E. Pek, and J. Starek. Large-scale, AST-based API-usage analysis of open-source Java projects. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC’11, pages 1317–1324, Mar. 2011. 170
806. T. K. Landauer. How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, 10:477–493, 1986. 50
807. R. M. Landers, J. B. Rebitzer, and L. J. Taylor. Rat race reduce: Adverse selection in the determination of work hours in law firms. *The American Economic Review*, 86(3):329–348, June 1996. 80
808. D. Landman, A. Serebrenik, E. Bouwers, and J. J. Vinju. Empirical analysis of the relationship between CC and SLOC in a large corpus of Java methods and C functions. *Journal of Software: Evolution and Process*, 28(7):589–618, July 2016. 139, 151, 164
809. D. Landy, D. Brookes, and R. Smout. Abstract numeric relations and the visual structure of algebra. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(5):1404–1418, Sept. 2014. 33
810. D. Landy, A. Charlesworth, and E. Ottmar. Categories of large numbers in line estimation. *Cognitive Science*, 41(2):326–353, Mar. 2017. 40
811. D. Landy and R. L. Goldstone. Proximity and precedence in arithmetic. *The Quarterly Journal of Experimental Psychology*, 63(10):1953–1968, Oct. 2010. 174
812. E. J. Langer. The illusion of control. *Journal of Personality and Social Psychology*, 32(2):311–328, 1975. 48
813. R. N. Langlois. External economies and economic progress: The case of the microcomputer industry. *The Business History Review*, 66(1):1–50, 1992. 78
814. LANL. Lanl failure data. <http://institute.lanl.gov/data/lanldata.shtml>, 2006. 143
815. L. Lapointe and S. Rivard. A multilevel model of resistance to information technology implementation. *MIS Quarterly*, 29(3):461–492, Sept. 2005. 102
816. I. Larkin. The cost of high-powered incentives: Employee gaming in enterprise software sales. Technical Report 13-073, Harvard Business School, Feb. 2013. 63
817. C. Larman and V. R. Basili. Iterative and incremental development: A brief history. *Computer*, 36(6):47–56, June 2003. 113, 114
818. J. Larres. Performance variance evaluation on Mozilla Firefox. Thesis (m.s.), Victoria University of Wellington, May 2012. 343
819. R. Latorre. Effects of developer experience on learning and applying unit test-driven development. *IEEE Transactions on Software Engineering*, 40(4):381–395, Apr. 2014. 26
820. S. Laumer, C. Maier, A. Eckhardt, and T. Weitzel. Work routines as an object of resistance during information systems implementations: theoretical foundation and empirical evidence. *European Journal of Information Systems*, 25(4):317–343, July 2016. 102
821. L. Lauterbach. Development of N-version software samples for an experiment in software fault tolerance. NASA Contractor Report 178363, Software Research and Development Center for Digital Systems Research, Sept. 1987. 109, 138
822. E. Lazear and M. Gibbs. *Personnel Economics for Managers*. John Wiley & Sons, Inc, second edition, 2007. 81
823. C. Lebiere. *The Dynamics of Cognition: An ACT-R Model of Cognitive Arithmetic*. PhD thesis, Carnegie Mellon University, Nov. 1998. 40
824. P. L’Ecuyer and R. Simard. TestU01: A C library for empirical testing of random number generators. *ACM Transactions on Mathematical Software*, 33(4):1–22, Aug. 2007. 141

825. A. L. Lederer and J. Prasad. Causes of inaccurate software development cost estimates. *Journal of Systems and Software*, 31(2):125–134, Nov. 1995. 105
826. B. C. Lee and D. M. Brooks. Regression modeling strategies for microarchitecture performance and power prediction. Technical Report TR-08-06, Division of Engineering and Applied Sciences, Harvard University, Mar 2006. 281, 324
827. D. Lee, Y. Kim, G. Pekhimenko, S. Khan, V. Seshadri, K. Chang, and O. Mutlu. Adaptive-latency DRAM: Optimizing DRAM timing for the common-case. In *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, HPCA, pages 489–501, Feb. 2015. 341
828. G. Leech, P. Rayson, and A. Wilson. *Word Frequencies in Written and Spoken English*. Pearson Education, 2001. 38, 164
829. L. Lefebvre and N. J. Boogert. Avian social learning. In M. D. Breed and J. Moore, editors, *Encyclopedia of Animal Behavior: volume 1*, pages 124–130. Oxford: Academic Press, July 2010. 71
830. J.-A. LeFevre and J. Liu. The role of experience in numerical skill: Multiplication performance in adults from Canada and China. *Mathematical Cognition*, 3(1):31–62, 1997. 41
831. G. E. Legge, T. A. Hooven, T. S. Klitz, J. S. Mansfield, and B. S. Tjan. Mr. Chips 2002: New insights from an ideal-observer model of reading. *Vision Research*, 42(18):2219–2234, Aug. 2002. 34
832. D. R. Lehman, R. O. Lempert, and R. E. Nisbett. The effects of graduate training on reasoning. *American Psychologist*, 43(6):431–442, 1988. 32
833. L. Lehmann, K. Aoki, and M. W. Feldman. On the number of independent cultural traits carried by individuals and populations. *Philosophical Transactions of The Royal Society B*, 366(1563):424–435, Feb. 2011. 71
834. P. Lemaire and M. Fayol. When plausibility judgments supersede fact retrieval: The example of the odd-even effect on product verification. *Memory & Cognition*, 23(1):34–48, Feb. 1995. 41
835. P. Lennie. The cost of cortical computation. *Current Biology*, 13(6):493–497, Mar. 2003. 50
836. K. Lerman, X. Yan, and X.-Z. Wu. The "majority illusion" in social networks. *PLoS ONE*, 11(2):e0147617, Feb. 2016. 68
837. B. W. Leverett, R. G. G. Cattell, S. O. Hobbs, J. M. Newcomer, A. H. Reiner, B. R. Schatz, and W. A. Wulf. An overview of the production-quality compiler-compiler project. Technical Report CMU-CS-79-105, Carnegie-Mellon University, Feb. 1979. 150
838. P. Lewicki, T. Hill, and E. Bizot. Acquisition of procedural knowledge about a pattern of stimuli that cannot be articulated. *Cognitive Psychology*, 20(1):24–37, Jan. 1988. 163
839. G. Lewis and P. Bajari. Incentives and adaptation: Evidence from highway procurement in Minnesota. Working Paper 17647, National Bureau of Economic Research, Dec. 2011. 108
840. X. Li. *Soft Error Modeling and Analysis for Microprocessors*. PhD thesis, University of Illinois at Urbana-Champaign, May 2008. 142
841. Z. Li, S. Lu, S. Myagmar, and Y. Zhou. CP-Miner: Finding copy-paste and related bugs in large-scale software code. *IEEE Transactions on Software Engineering*, 32(3):176–192, Mar. 2006. 95
842. Y. Liang, Y. Zhang, A. Sivasubramaniam, R. K. Sahoo, J. Moreira, and M. Gupta. Filtering failure logs for a BlueGene/L prototype. In *Proceedings of the 2005 International Conference on Dependable Systems and Networks*, DSN'05, pages 476–485, June 2005. 371
843. S. Lichtenstein and B. Fishhoff. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20:159–183, 1977. 48
844. Y. Lichtenstein and A. McDonnell. Pricing software development services. In *ECIS 2003 Proceedings*, 2003. 111
845. G. A. Liebchen and M. Shepperd. Data sets and data quality in software engineering. In *Proceedings of the 4th international workshop on Predictor models in software engineering*, PROMISE'08, pages 39–44, May 2008. 363
846. J. Liebig, S. Apel, C. Lengauer, C. Kästner, and M. Schulze. An analysis of the variability in forty preprocessor-based software product lines. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering-Volume 1*, ICSE'10, pages 105–114, May 2010. 152
847. J. H. Lienhard. *How Invention Begins: Echoes of Old Voices in the Rise of New Machines*. Oxford University Press, 2006. 68
848. B. P. Lientz, E. B. Swanson, and G. E. Tompkins. Characteristics of application software maintenance. *Communications of the ACM*, 21(6):466–471, June 1978. 85
849. J. S. Light. When computers were women. *Technology and Culture*, 40(3):455–483, July 1999. 72
850. S. L. Lim. *Social Networks and Collaborative Filtering for Large-Scale Requirements Elicitation*. PhD thesis, School of Computer Science and Engineering, University of New South Wales, Aug. 2010. 112, 117, 118
851. D.-Y. Lin and I. Neamtiu. Collateral evolution of applications and databases. In *Proceedings of the joint international and annual ERCIM workshops on Principles of software evolution and Software Evolution (Evol) workshops*, IWPSE-Evol'09, pages 31–40, Aug. 2009. 96
852. S. Lin. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570, Sept. 2010. 346
853. M. Linares-Vásquez, G. Bavota, C. Bernal-Cárdenas, M. Di Penta, R. Oliveto, and D. Poshyvanyk. API change and fault proneness: A threat to the success of Android apps. In *Proceedings of the 9th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2013, pages 477–487, Aug. 2013. 130
854. K. R. Linberg. Software developer perceptions about software project failure: a case study. *The Journal of Systems and Software*, 49(2–3):177–192, Dec. 1999. 102
855. K. Lind and R. Heldal. A practical approach to size estimation of embedded software components. *IEEE Transaction on Software Engineering*, 38(5):993–1007, Sept.-Oct. 2012. 109
856. R. Lister, E. S. Adams, S. Fitzgerald, W. Fone, J. Hamer, M. Lindholm, R. McCartney, J. E. Moström, K. Sanders, O. Seppälä, B. Simon, and L. Thomas. A multi-national study of reading and tracing skills in novice programmers. *ACM SIGCSE Bulletin*, 36(4):119–150, Dec. 2004. 320
857. T. Little. Schedule estimation and uncertainty surrounding the cone of uncertainty. *IEEE Software*, 23(3):48–54, May 2006. 115, 116
858. D. C. Littman, J. Pinto, S. Letovsky, and E. Soloway. Mental models and software maintenance. In E. Soloway and S. Iyengar, editors, *Empirical Studies of Programmers*, chapter 6, pages 80–98. Ablex Publishing Corporation, 1986. 154
859. S. Livieri, Y. Higo, M. Matsushita, and K. Inoue. Analysis of the Linux kernel evolution using code clone coverage. In *Fourth International Workshop on Mining Software Repositories*, MSR'07, pages 22–25, May 2007. 93
860. G. D. Logan. Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5):883–914, 1992. 25
861. C. V. Lopes, P. Maj, P. Martins, V. Saini, D. Yang, J. Zitny, H. Sajnani, and J. Vitek. DéjàVu: A map of code duplicates on GitHub. In *Conference on Object-Oriented Programming Systems, Languages, and Applications*, OOPSLA'17, page ???, Oct. 2017. 165, 166
862. P. Louridas, D. Spinellis, and V. Vlachos. Power laws in software. *ACM Transaction on Software Engineering and Methodology*, 18(1):1–26, Sept. 2008. 276
863. L. Lu, A. C. Arpacı-Dusseau, R. H. Arpacı-Dusseau, and S. Lu. A study of Linux file system evolution. In *11th USENIX Conference on File and Storage Technologies*, FAST'13, pages 31–44, Feb. 2013. 166, 181
864. J. D. Luente. *On the Viability of Quantitative Assessment Methods in Software Engineering and Software Services*. PhD thesis, School of Engineering and Computer Science, University of Denver, Jan. 2015. 131
865. L. Lucia. *Ranking-Based Approaches for Localizing Faults*. PhD thesis, Singapore Management University, June 2014. 140
866. L. Lucia, D. Lo, L. Jiang, F. Thung, and A. Budi. Extended comprehensive study of association measures for fault localization. *Journal of Software: Evolution and Process*, 26(2):172–219, Feb. 2014. 139
867. K. M. Lui and K. C. C. Chan. Pair programming productivity: Novice-novice vs. expert-expert. *International Journal of Human-Computer Studies*, 64(9):915–925, Sept. 2006. 25

868. A. C.-A. Luís M.A. Bettencourt and, D. I. Kaiserd, and C. Castillo-Chávez. The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models. *Physica A*, 364:513–536, May 2006. 72
869. P. Lukowicz, E. A. Heinz, L. Prechelt, and W. F. Tichy. Experimental evaluation in computer science: A quantitative study. Technical Report 17/94, University of Karlsruhe, Germany, Aug. 1994. 5
870. Y. Luo, S. Govindan, B. Sharma, M. Santaniello, J. Meza, A. Kansal, J. Liu, B. Khessib, K. Vaid, and O. Mutlu. Characterizing application memory error vulnerability to optimize datacenter cost via heterogeneous-reliability memory. In *Proceedings of the 2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 467–478, June 2014. 142
871. A. K. Luria. Towards the problem of the historical nature of psychological processes. *International Journal of Psychology*, 6(4):259–272, 1971. 17, 35
872. A. R. Luria. *The Mind of a Mnemonist*. Penguin Education, 1975. 23
873. B. Luthiger and C. Jungwirth. Pervasive fun. *First Monday*, 12(1), Jan. 2007. 19
874. R. Lutz. Analyzing software requirements errors in safety-critical, embedded systems. TR 92–27, Department of Computer Science, Iowa State University of Science and Technology, Aug. 1992. 138
875. W. J. Lynn III. A new approach for delivering information technology capabilities in the department of defense. Report to congress, Office of the Secretary of Defense, Nov. 2010. 113
876. W. Ma, L. Chen, X. Zhang, Y. Zhou, and B. Xu. How do developers fix cross-project correlated bugs? A case study on the GitHub scientific Python ecosystem. In *IEEE/ACM 39th International Conference on Software Engineering*, ICSE’17, pages 381–392, May 2017. 127
877. W. Ma, J.-C. S. Liu, and A. Forin. Design and testing of a cpu emulator. Technical Report MSR-TR-2009-155, Microsoft Research, Aug. 2009. 141
878. N. Macdonald. Computing services survey. *Computers and Automation*, 7(7):9–12, July 1958. 77
879. N. Macdonald. *Computer Census 1962-74*. Computers and People, 1974. 76
880. J. N. MacGregor. Short-term memory capacity: Limitation or optimization? *Psychological Review*, 94(1):107–108, 1987. 24
881. A. Machiry, R. Tahiliani, and M. Naik. Dynodroid: An input generation system for Android apps. In *Proceedings of the 9th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE’13, pages 224–234, Aug. 2013. 307, 308
882. C. E. Mackenzie. *Coded Character Sets, History and Development*. Addison-Wesley, 1980. 72
883. D. MacKenzie. Computer-related accidental death: an empirical exploration. *Science and Public Policy*, 21(4):233–248, Aug. 1994. 128
884. R. J. Madachy. *Software Process Dynamics*. John Wiley & Sons, Inc, 2008. 107
885. A. Maddison. Business cycles, long waves and phases of capital development. In A. Maddison, editor, *Dynamic Forces in Capitalist Development: A Long-run Comparative View*, chapter 4, page ??? Oxford University Press, Oct. 1991. 5
886. W. T. Maddox and C. J. Bohil. Costs and benefits in perceptual categorization. *Memory & Cognition*, 28:597–615, 2000. 27, 311
887. M. Magidin and E. Viso. On the experiments in algorithm dynamics. Technical Report UAMR0853, Universidad Autónoma Metropolitana-Iztapalapa, México, Oct. 1976. 164
888. T. Maillart, M. Zhao, J. Grossklags, and J. Chuang. Given enough eyeballs, all bugs are shallow? Revisiting Eric Raymond with bug bounty programs. *Journal of Cybersecurity*, ???(??):1–10, Apr. 2017. 71, 129
889. V. N. Makarov. SPEC benchmark page. <http://vmakarov.fedorapeople.org/spec/index.html>, July 2014. 343, 344
890. B. A. Malloy and J. F. Power. Quantifying the transition from Python 2 to 3: An empirical study of Python applications. In *International Symposium on Empirical Software Engineering and Measurement*, ESEM’17, pages 314–323, Dec. 2017. 166
891. M. Mangel and F. J. Samaniego. Abraham Wald’s work on aircraft survivability. *Journal of the American Statistical Association*, 79(386):259–267, June 1984. 221
892. M. V. Mäntylä and J. Itkonen. How are software defects found? The role of implicit defect detection, individual responsibility, documents, and knowledge. *Information and Software Technology*, 56(12):1597–1612, Dec. 2014. 147
893. A. Marchand. The power of an installed base to combat lifecycle decline: The case of video games. *International Journal of Research in Marketing*, 33(1):140–154, Mar. 2016. 63
894. J. N. Marewski and L. J. Schooler. Cognitive niches: An ecological model of strategy selection. *Psychological Review*, 118(3):292–437, 2011. 44
895. B. H. Margolin, R. P. Parmelee, and M. Schatzoff. Analysis of free-storage algorithms. *IBM Systems Journal*, 10(4):283–304, 1971. 167
896. P. Marinescu, P. Hosek, and C. Cedar. COVRIG: A framework for the analysis of code, test, and coverage evolution in real software. In *Proceedings of the 2014 International Symposium on Software Testing and Analysis*, ISSTA’14, pages 93–104, July 2014. 146
897. F. Marotta-Wurgler. What’s in a standard form contract? An empirical analysis of software license agreements. *Journal of Empirical Legal Studies*, 7(4):677–713, Dec. 2007. 111
898. F. Marotta-Wurgler. Will increased disclosure help? Evaluating the recommendations of the ALI’s “principles of the law of software contracts”. *University of Chicago Law Review*, 78(1), 2011. 92
899. D. H. Martin, J. R. Cordy, B. Adams, and G. Antoniol. Make it simple—An empirical analysis of GNU make feature use in open source projects. In *IEEE 23rd International Conference on Program Comprehension*, pages 207–217, May 2015. 172
900. F. Martineau. PNFG: A framework for computer game narrative analysis. Thesis (m.s.), School of Computer Science, McGill University, June 2006. 168
901. A. G. Martínez. *Chaos Monkeys: Inside The Silicon Valley Money Machine*. Ebury Press, 2016. 79
902. E. Masanet, A. Shehabi, J. Liang, L. Ramakrishnan, X. Ma, V. Hendrix, B. Walker, and P. Mantha. The energy efficiency potential of cloud-based software: A U.S. case study. LBNL Paper LBNL-6298E, Lawrence Berkeley National Laboratory, June 2013. 78
903. F. Massacci, S. Neuhaus, and V. H. Nguyen. After-life vulnerabilities: A study on Firefox evolution, its vulnerabilities, and fixes. In *Proceedings of the Third international conference on Engineering secure software and systems*, ESSoS’11, pages 195–208, Feb. 2011. 135, 136
904. E. Matias, I. S. MacKenzie, and W. Buxton. One-handed touch-typing on a QWERTY keyboard. *Human-Computer Interaction*, 11:1–27, 1996. 42
905. C. Mayer, S. Hanenberg, R. Robbes, É. Tanter, and A. Stefik. An empirical study of the influence of static type systems on the usability of undocumented software. In *Proceedings of the ACM international conference on Object oriented programming systems languages and applications*, OOPSLA’12, pages 683–702, Oct. 2012. 153
906. D. Mazinanian, A. Ketkar, N. Tsantalis, and D. Dig. Understanding the use of lambda expressions in Java. In *Proceedings of the ACM on Programming Languages*, OOPSLA’17, page 85, Oct. 2017. 167
907. A. Mazouz. *An Empirical Study of Program Performance of OpenMP Applications on Multicore Platforms*. PhD thesis, Université de Versailles-Saint Quentin en Yvelines, Dec. 2013. 342
908. D. F. McAllister and M. A. Vouk. Experiments in fault tolerant software reliability. Technical Report 5 – NAG-1-667, North Carolina State University, Apr. 1989. 109, 147, 148
909. T. J. McCabe. A complexity measure. *IEEE Transactions on Software Engineering*, SE-2(4):308–320, Dec. 1976. 164
910. J. C. McCallum. Historical cost of computer memory and storage. <http://www.jcmit.com>, July 2016. 1
911. S. McCloud. *Understanding Comics*. HarperPerennial, 1993. 192
912. S. McConnell. *Code Complete*. Microsoft Press, 1993. 152
913. M. H. McCormack. *The Terrible Truth about Lawyers*. Beech Tree books, William Morrow, 1987. 110
914. M. McCracken, V. Almstrum, D. Diaz, M. Guzdial, D. Hagan, Y. B.-D. Kolikant, C. Laxer, L. Thomas, I. Utting, and T. Wilusz. A multi-national, multi-institutional study of assessment of programming skills of first-year CS students. *ACM SIGCSE Bulletin*, 33(4):125–180, June 2001. 320

915. R. R. McCrae and P. T. Costa, Jr. Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. *Journal of Personality*, 57(1):17–40, Mar. 1989. 43
916. B. D. McCullough. Microsoft Excel's 'Not The Wichmann-Hill' random number generator. *Computational Statistics and Data Analysis*, 52:4587–4593, 2008. 141
917. B. D. McCullough and D. A. Heiser. On the accuracy of statistical procedures in Microsoft Excel 2007. *Computational Statistics and Data Analysis*, 52:4570–4578, 2008. 11
918. R. McElreath and R. Boyd. *Mathematical Models of Social Evolution: A Guide for the Perplexed*. The University of Chicago Press, 2008. 71
919. D. McFadden. Rationality for economists? *Journal of Risk and Uncertainty*, 19:73–105, 1999. 44
920. K. B. McKeithen, J. S. Reitman, H. H. Ruster, and S. C. Hirtle. Knowledge organization and skill differences in computer programmers. *Cognitive Psychology*, 13(3):307–325, July 1981. 31
921. J. McManus and T. Wood-Harper. Understanding the sources of information systems project failure: A study in IS project failure. *Management Services*, 51(3):38–43, Aug. 2007. 104
922. D. S. McNamara and J. Magliano. Toward a comprehensive model of comprehension. In B. Ross, editor, *The Psychology of Learning and Motivation*, Vol. 51, chapter 9, pages 297–384. Academic Press, Nov. 2009. 158
923. T. P. McNamara, J. K. Hardy, and S. C. Hirtle. Subjective hierarchies in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2):211–227, 1989. 34, 35
924. J. McNerney, J. D. Farmer, S. Redner, and J. E. Trancik. Role of design complexity in technology improvement. *Proceedings of the National Academy of Sciences*, 108(22):9008–9013, May 2011. 76
925. D. L. McNicol. Influences on the timing and frequency of cancellations and truncations of major defense acquisition programs. IDA Paper P-8280, Institute for Defense Analyses, Mar. 2017. 104
926. I. McPhee. Customs' cargo management re-engineering project: Australian customs service. Audit Report No.24 2006-07, Australian National Audit Office, Aug. 2007. 102
927. J. H. McWhorter. The world's simplest grammars are creole grammars. *Linguistic Typology*, 5(2-3):125–166, 2001. 166
928. A. D. Meacham. *Data Processing Equipment Encyclopedia: Electronic Devices*, volume 2. Gille Associates, Inc., 1962. 336
929. F. Mechner. Probability relations within response sequences under ratio reinforcement. *Journal of the Experimental Analysis of Behavior*, 1(2):109–121, Apr. 1958. 16
930. F. Medeiros, C. Kästner, M. Ribeiro, S. Nadi, and R. Gheyi. The love/hate relationship with the C preprocessor: An interview study. In *Proceedings of the 29th European Conference on Object-Oriented Programming*, ECOOP'15, pages 495–518, July 2015. 152
931. M. Mehrara and T. Austin. Exploiting selective placement for low-cost memory protection. *ACM Transactions on Architecture and Code Optimization*, 5(3):1–24, Nov. 2008. 142
932. L. K. Melhus and R. E. Jensen. Measurement bias from address aliasing. In *Eleventh International Workshop on Automatic Performance Tuning*, iWAPT 2016, pages 1506–1515, May 2016. 340
933. D. Mendez, B. Baudry, and M. Monperrus. Empirical evidence of large-scale diversity in API usage of object-oriented software. In *International Conference on Source Code Analysis and Manipulation*, SCAM'2013, pages 43–52, Apr. 2013. 170
934. H. Mercier and D. Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34(2):57–111, Apr. 2011. 36
935. R. C. Merkle. Energy limits to the computational power of the human brain. *Foresight Update*, 6, Aug. 1989. 50
936. E. W. Merrow, L. McDonnel, and R. W. Argüden. Understanding the outcomes of megaprojects: A quantitative analysis of very large civilian projects. Report R-3560-PSSP, The RAND Corporation, Mar. 1988. 105
937. S. Mertens and C. Baethge. The virtues of correct citation. *Deutsches Ärzteblatt International*, 108(33):550–552, Apr. 2011. 124
938. A. Mesoudi. Cultural evolution: A review of theory, findings and controversies. *Evolutionary Biology*, 43(4):481–497, Dec. 2016. 71
939. T. Micceri. The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1):156–166, Apr. 1989. 219
940. German comments in LibreOffice. website, Apr. 2017. <https://people.gnome.org/~michael/data/2015-08-01-5.5-data.xls>. 73
941. S. E. Michalak, A. J. DuBois, C. B. Storlie, H. M. Quinn, W. N. Rust, D. H. DuBois, D. G. Modl, A. Manuzzato, and S. P. Blanchard. Assessment of the impact of cosmic-ray-induced neutrons on hardware in the Roadrunner supercomputer. *IEEE Transactions on Device and Materials Reliability*, 12(2):445–454, May 2012. 142
942. J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 14(6014):176–182, Jan. 2011. 91, 371
943. 2010 state of the computer book market. website, Feb. 2011. <http://radar.oreilly.com/2011/02/2010-book-market-1.html>. 90
944. S. Milgram. *Obedience to Authority*. McGraw-Hill, 1974. 47
945. S. Milgram, L. Bickman, and L. Berkowitz. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology*, 13(2):79–82, 1969. 72
946. A. Mili, S. F. Chmiel, R. Gottumukkala, and L. Zhang. Managing software reuse economics: An integrated ROI-based model. *Annals of Software Engineering*, 11(1):175–218, Nov. 2001. 95
947. K. Milis. *Success factors for ICT projects: Empirical research, utilizing qualitative and quantitative approaches (incl. Bayesian networks, Probabilistic feature models)*. PhD thesis, Toegepaste Economische Wetenschappen, Limburgs Universitair Centrum, Dec. 2002. 102
948. D. M. Miller. Application of halstead's timing model to predict the compilation time of Ada compilers. Thesis (m.s.), School of Engineering of the Air Force Institute of Technology, Air University, Dec. 1986. 151
949. D. R. Miller. Exponential order statistic models of software reliability growth. NASA Contractor Report 3909, NASA Langley Research Center, July 1985. 127, 133
950. G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97, 1956. 20, 322
951. G. A. Miller and S. Isard. Free recall of self-embedded English sentences. *Information and Control*, 7:292–303, 1964. 22
952. L. A. Miller. Natural language programming: Styles, strategies, and contrasts. *IBM Systems Journal*, 29(2):184–215, 1981. 154
953. S. J. Miller and M. J. Nigrini. Order statistics and Benford's law. *International Journal of Mathematics and Mathematical Sciences*, 2008, 2008. 372
954. W. R. Miller and M. Sanchez-Craig. How to have a high success rate in treatment: advice for evaluators of alcoholism programs. *Addiction*, 91(6):779–785, Apr. 1996. 319
955. S. MINAKAWA, T. HIRATA, K. MASAME, H. OKADA, and K. MARUYAMA. A psychological analysis on the meaning of "reliance". *Tohoku Psychologica Folia*, 46(1-4):111–117, Apr. 1987. 127
956. C. H. Mireles. *Marketing Modeling for New Products*. PhD thesis, Erasmus University Rotterdam, June 2010. 63
957. MISRA. *MISRA-C:2004 Guidelines for the Use of the C Language in Vehicle Based Software*. Motor Industry Research Association, 2004. 128, 152
958. MISRA. *MISRA-C++:2008 Guidelines for the Use of the C++ Language in Critical Systems*. Motor Industry Research Association, June 2008. 128
959. K. Mitropoulou. *Performance Optimizations for Compiler-based Error Detection*. PhD thesis, School of Informatics, University of Edinburgh, Oct. 2014. 142
960. S. Mittal. A survey of architectural techniques for managing process variation. *ACM Computing Surveys*, 48(4), May 2016. 337
961. M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003. 203
962. M. Mitzenmacher. Dynamic models for file sizes and double Pareto distributions. *Internet Mathematics*, 1(3):305–333, Apr. 2004. 192, 211

963. A. Mockus and L. G. Votta. Identifying reasons for software changes using historic databases. In *Proceedings of the International Conference on Software Maintenance, ICSM'00*, page ???, Oct. 2000. 85
964. A. Mockus and D. M. Weiss. Predicting risk of software changes. *Bell Labs Technical Journal*, 5(2), Apr.-June 2000. 25
965. S. N. Mohanty. Software cost estimation: Present and future. *Software—Practice and Experience*, 11(2):103–121, Feb. 1981. 106
966. T. Moher and G. M. Schneider. Methods for improving controlled experimentation in software engineering. In *Proceedings of the 5th International Conference on Software Engineering*, pages 224–233. IEEE Computer Society, Mar. 1981. 320
967. K. Moløkken-Østvold and K. M. Furulund. The relationship between customer collaboration and software project overruns. In *2007 Agile Conference, AGILE*, pages 72–83, Aug. 2007. 267
968. K. Moløkken-Østvold, M. Jørgensen, S. S. Tanilkan, H. Gallis, A. C. Lien, and S. E. Hove. A survey on software estimation in the norwegian industry. In *Proceedings 10th International Symposium on Software Metrics*, pages 208–219, Sept. 2004. 108
969. G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, Apr. 1965. 4
970. T. Moscibroda and R. Oshman. Resilience of mutual exclusion algorithms to transient memory faults. In *Proceedings of the 30th annual ACM SIGACT-SIGOPS symposium on Principles of distributed computing, PODC'11*, pages 69–78, June 2011. 142
971. J. F. Motz. In re microsoft corporation antitrust litigation \* sun microsystems, inc. v. microsoft corporation \* mdl 1332 \* civil no. jfm-02-2739. Opinion, UNITED STATES DISTRICT COURT FOR THE DISTRICT OF MARYLAND, 2002. 70
972. Y. Moy and A. Wallenburg. Tokeneer: Beyond formal program verification. In *Proceedings of the 5th International Congress on Embedded Real Time Software and Systems, ERTS<sup>2</sup> 2010*, May 2010. 124
973. S. T. Mueller and A. Krawitz. Reconsidering the two-second decay hypothesis in verbal working memory. *Journal of Mathematical Psychology*, 53(1):14–25, Feb. 2009. 21
974. S. T. Mueller and C. T. Weidemann. Alphabetic letter identification: Effects of perceptability, similarity, and bias. *Acta Psychologica*, 139(1):19–37, Jan. 2012. 42, 162
975. D. Mulcahy, B. Weeks, and H. S. Bradley. "WE HAVE MET THE ENEMY... AND HE IS US" Lessons from twenty years of the Kauffman Foundation's investments in venture capital funds and the triumph of hope over experience. Report, Ewing Marion Kauffman Foundation, May 2012. 80
976. C. W. Mulford and S. Misra. Capitalization of software development costs: Accounting practices in the software industry, 2014 and 2015. Technical report, Georgia Tech College of Management, Jan. 2016. 53, 58, 60
977. C. W. Mulford and J. Roberts. Capitalization of software development costs: A survey of accounting practices in the software industry. Technical report, Georgia Tech College of Management, May 2006. 60
978. M. M. Müller and A. Höfer. The effect of experience on the test-driven development process. *Empirical Software Engineering*, 12(6):593–615, 2007. 178, 320, 365
979. J. Mulligan and B. Patrovsky. *Developing Online Games: An Insider's Guide*. New Riders Publishing, 2003. 111
980. D. Muna, M. Alexander, A. Allen, R. Ashley, D. Asmus, R. Azollini, M. Bannister, R. Beaton, A. Benson, G. B. Berriman, M. Bilicki, P. Boyce, J. Bridge, J. Cami, E. Cangi, X. Chen, N. Christiny, C. Clark, M. Collins, J. Comparat, N. Cook, D. Croton, I. D. Davids, É. Depagne, J. Donor, L. A. dos Santos, S. Douglas, A. Du, M. Durbin, D. Erb, D. Faes, J. G. Fernández-Trincado, A. Foley, S. Fotopoulos, S. Frimann, P. Frinchaboy, R. Garcia-Dias, A. Gawryszczak, E. George, S. Gonzalez, K. Gordon, N. Gorgone, C. Gosmeyer, K. Grasha, P. Greenfield, R. Grellmann, J. Guillouchon, M. Gurwell, M. Haas, A. Hagen, D. Haggard, T. Haines, P. Hall, W. Hellwing, E. C. Herenz, S. Hinton, R. Hlozek, J. Hoffman, D. Holman, B. W. Holwerda, A. Horton, C. Hummels, D. Jacobs, J. J. Jensen, D. Jones, A. Karick, L. Kelley, M. Kenworthy, B. Kitchener, D. Klaes, S. Kohn, P. Konorski, C. Krawczyk, K. Kuehn, T. Kuutma, M. T. Lam, R. Lane, J. Liske, D. Lopez-Camara, K. Mack, S. Mangham, Q. Mao, D. J. E. Marsh, C. Mateu, L. Maurin, J. McCormac, I. Momcheva, H. Monteiro, M. Mueller, R. Munoz, R. Naidu, N. Nelson, C. Nitschelm, C. North, J. Nunez-Iglesias, S. Ogaz, R. Owen, J. Parejko, V. Patrício, J. Pepper, M. Perrin, T. Pickering, J. Piscionere, R. Pogge, R. Poleski, A. Pourtsidou, A. M. Price-Whelan, M. L. Rawls, S. Read, G. Rees, H. Rein, T. Rice, S. Riemer-Sørensen, N. Rusomarov, S. F. Sanchez, M. Santander-García, G. Sarid, W. Schoenell, A. Scholz, R. L. Schuhmann, W. Schuster, P. Scicluna, M. Seidel, L. Shao, P. Sharma, A. Shulevski, D. Shupe, C. Sifón, B. Simmons, M. Sinha, I. Skillen, B. Soergel, T. Spriggs, S. Srinivasan, A. Stevens, O. Streicher, E. Suchtya, J. Tan, O. G. Telford, R. Thomas, C. Tonini, G. Tremblay, S. Tuttle, T. Urrutia, S. Vaughan, M. Verdugo, A. Wagner, J. Walawender, A. Wetzel, K. Willett, P. K. G. Williams, G. Yang, G. Zhu, and A. Zonca. The Astropy problem. In *eprint arXiv:astro-ph.IM/1610.03159*, Oct. 2016. 81, 103
981. B. B. Murdoch, Jr. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488, 1962. 24
982. P. Murrell. *R Graphics*. Chapman & Hall/CRC, 1st edition, 2006. 189
983. L. H. Mutuel. Single event effects mitigation techniques report. Final Report DOT/FAA/TC-15/62, U.S. Department of Transportation, Federal Aviation Administration, Feb. 2016. 142
984. G. J. Myers. A controlled experiment in program testing and code walkthroughs/inspections. *Communications of the ACM*, 21(9):760–768, Sept. 1978. 144
985. T. Mytkowicz, A. Diwan, M. Hauswirth, and P. F. Sweeney. We have it easy, but do we have it right? In *International Symposium on Parallel and Distributed Processing, IPDPS 2008*, pages 1–7, Apr. 2008. 340, 342
986. M. Nagappan, R. Robbes, Y. Kamei, É. Tanter, S. McIntosh, A. Mockus, and A. E. Hassan. An empirical study of goto in C code from GitHub repositories. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015*, pages 404–414, Aug.-Sept. 2015. 169
987. M. Nagappan, T. Zimmermann, and C. Bird. Diversity in software engineering research. In *Proceedings of the 9th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2013*, pages 466–476, Aug. 2013. 222
988. N. Nagappan, A. Zeller, T. Zimmermann, K. Herzig, and B. Murphy. Change bursts as defect predictors. In *IEEE 21st International Symposium on Software Reliability Engineering, ISSRE*, pages 309–318, Nov. 2010. 272
989. P. M. Nagel, F. W. Scholz, and J. A. Skrivan. Software reliability: Additional investigations into modeling with replicated experiments. NASA Contractor Report 172378, Boeing Computer Services Company, Space and Military Applications Division, June 1984. 132
990. P. M. Nagel and J. A. Skrivan. Software reliability: Repetitive run experimentation and modeling. NASA Contractor Report 165836, Boeing Computer Services Company, Space and Military Applications Division, Feb. 1982. 132
991. T. Nagle, J. Hogan, and J. Zale. *The Strategy and Tactics of Pricing*. Pearson, fifth edition, 2015. 61
992. J. Nandakumar and D. E. Avison. The fiction of methodological development: a field study of information systems development. *Information Technology & People*, 12(2):176–191, Feb. 1999. 114
993. S. Nanz and C. A. Furia. A comparative study of programming languages in Rosetta Code. In *eprint arXiv:cs.SE/1409.0252v1*, Aug. 2014. 153
994. M. B. Nathanson. Desperately seeking mathematical truth. *Notices of the AMS*, 55(7):773–773, Aug. 2008. 124
995. P. Naur and B. Randell. Software engineering report on a conference sponsored by the NATO science committee. Technical report, NATO, Jan. 1969. 5, 70
996. D. J. Navarro and A. F. Perfors. Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, 118(1):120–134, Jan. 2011. 52
997. I. Neamtiu, J. S. Foster, and M. Hicks. Understanding source code evolution using abstract syntax tree matching. In *Proceedings of the 2005 International Workshop on Mining Software Repositories*, pages 1–5, May 2005. 171
998. D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The university of South Florida word association, rhyme and word fragment norms. [w3.usf.edu/FreeAssociation](http://w3.usf.edu/FreeAssociation), 1998. 161

999. E. A. Nelson. Management handbook for the estimation of computer programming costs. Technical Documentary Report ESD-TDR-67-66, United States Air Force, L. G. Hanscom Field, Bedford, Massachusetts, Oct. 1966. [106](#)
1000. D. A. Nembhard and N. Osothsilp. An empirical comparison of forgetting models. *IEEE Transactions on Engineering Management*, 48(3):283–291, Aug. 2001. [76](#)
1001. R. E. NeSmith II. A study of software maintenance costs of Air Force large scale computer systems. Thesis (m.s.), School of Systems and Logistics, Air Force Institute of Technology, Air University, Sept. 1986. [77](#), [260](#)
1002. B. New, L. Ferrand, C. Pallier, and M. Brysbaert. Reexamining the word length effect in visual word recognition: New evidence from the English lexicon project. *Psychonomic Bulletin & Review*, 13(1):45–52, Feb. 2006. [161](#)
1003. A. Newell. *Unified Theories of Cognition*. Harvard University Press, 1991. [16](#)
1004. A. Newell and P. S. Rosenbloom. Mechanisms of skill acquisition and the power law of practice. Technical report, Carnegie Mellon University, Aug. 1982. [24](#)
1005. G. Nezlek and G. DeHondt. An empirical investigation of gender wage differences in information systems occupations: 1991–2008. In *Proceedings of the 43rd Hawaii International Conference on System Sciences–2010*, pages 4059–4068, Jan. 2010. [83](#)
1006. T. H. D. Nguyen, B. Adams, and A. E. Hassan. A case study of bias in bug-fix datasets. In *17th Working Conference on Reverse Engineering*, WCRE, pages 259–268, Oct. 2010. [127](#)
1007. V. H. Nguyen and F. Massacci. The (un)reliability of NVD vulnerable versions data: an empirical experiment on google chrome vulnerabilities. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, ASIA CCS’13, pages 493–498, May 2013. [127](#)
1008. J. Nielsen and T. K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERCHI ’93 conference on Human factors in computing systems*, INTERCHI’93, pages 206–213, Apr. 1993. [137](#)
1009. E. B. Nightingale, J. R. Douceur, and V. Orgovan. Cycles, cells and platters: An empirical analysis of hardware failures on a million consumer PCs. In *Proceedings of the sixth conference on Computer systems*, EuroSys’11, pages 343–356, Apr. 2011. [334](#)
1010. M. J. Nigrini and S. J. Miller. Data diagnostics using second order tests of Benford’s law. *Auditing: A Journal of Practice and Theory*, 28(2):305–324, June 2009. [372](#)
1011. D. E. Nikonorov and I. A. Young. Overview of beyond-CMOS devices and a uniform methodology for their benchmarking. In *eprint arXiv:cond-mat.mes-hall/1302.0244*, Feb. 2013. [339](#)
1012. J. Ninio and K. A. Stevens. Variations on the Hermann grid: an extinction illusion. *Perception*, 29(10):1209–1217, Oct. 2000. [159](#)
1013. R. E. Nisbett, D. H. Krantz, C. Jepson, and Z. Kunda. The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90(4):339–363, 1983. [27](#)
1014. NIST???. National vulnerability database. <https://nvd.nist.gov>, Dec. 2014. [126](#), [366](#)
1015. G. P.-C. no and D. M. German. Software patents: A replication study. In *Proceedings of the 11th International Symposium on Open Collaboration*, OpenSym’15, pages 5:1–5:4, Aug. 2015. [8](#)
1016. S. Nørby. Why forget? On the adaptive value of memory loss. *Perspectives on Psychological Science*, 10(5):551–578, Sept. 2015. [23](#)
1017. P. V. Norden. Resource usage and network planning techniques. In B. V. Dean, editor, *Operations Research in Research and Development*, chapter 5, pages 149–169. John Wiley & Sons, Inc, 1963. [107](#)
1018. W. D. Nordhaus. The progress of computing. Cowles Foundation Discussion Paper No. 1324, Yale University, Sept. 2001. [1](#)
1019. J. A. Norton and F. M. Bass. A diffusion theory model of adoption and substitution for successive generations of high-technology products. *Management Science*, 33(9):1069–1086, Sept. 1987. [63](#)
1020. M. A. Nowak. *Evolutionary Dynamics: Exploring the Equations of Life*. The Belknap press of Harvard University press, 2006. [97](#)
1021. D. Nowroth, I. Polian, and B. Becker. A study of cognitive resilience in a JPEG compressor. In *IEEE International Dependable Systems and Networks With FTCS and DCC*, DSN 2008, pages 32–41, June 2008. [141](#)
1022. H.-C. Nuerk, G. Wood, and K. Willmes. The universal snarc effect: The association between number magnitude and space is amodal. *Experimental Psychology*, 52(3):187–194, 2005. [18](#)
1023. R. E. Núñez. No innate number line in the human brain. *Journal of Cross-Cultural Psychology*, 42(4):651–668, 2011. [40](#)
1024. J. M. Nuttin Jr. Affective consequence of mere ownership: The name letter effect in twelve European languages. *European Journal of Social Psychology*, 17:381–402, 1987. [73](#)
1025. NVIDIA. *CUDA CUBLAS Library*. NVIDIA Corporation, CA, USA, 3.1 edition, Aug. 2010. [321](#)
1026. M. Oaksford and N. Chater. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4):608–631, 1994. [36](#)
1027. K. Oberauer, H.-M. Süß, O. Wilhelm, and W. W. Wittmann. The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, 31:167–193, 2003. [21](#)
1028. O. O. Odeh, A. M. Featherstone, and J. S. Bergtold. Reliability of statistical software. *American Journal of Agricultural Economics*, 92(5):1472–1489, Sept. 2010. [11](#)
1029. OECD. *OECD Digital Economy Outlook 2015*. OECD Publishing, 2015. [70](#)
1030. P. Oladimeji. Devices, errors and improving interaction design-A case study using an infusion pump. Thesis (m.res.), Department of Computer Science, Swansea University, Oct. 2008. [212](#)
1031. A. Oliner and J. Stearley. What supercomputers say: A study of five system logs. In *37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, DSN’07, pages 575–584, June 2007. [143](#)
1032. P. Oliver. Experiences in building and using compiler validation systems. In R. Merwin and J. Zanca, editors, *AFIPS Conference Proceedings*, pages 1051–1057. AFIPS Press, June 1979. [146](#)
1033. T. Open Group. The Austin common standards revision group. <http://austingroupbugs.net>, July 2017. [138](#)
1034. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, Aug. 2015. [7](#)
1035. OpenCorporates. UK registered company data. <https://opencorporates.com/>, Mar. 2015. [79](#)
1036. OpenRefine. website, Oct. 2014. <http://openrefine.org>. [364](#)
1037. OpenSignal. Android fragmentation visualized (august 2015). Technical Report ???, OpenSignal, Aug. 2015. [189](#), [190](#), [365](#)
1038. N. Osaka. Eye fixation and saccade during kana and kanji text reading: Comparison of English and Japanese text processing. *Bulletin of the Psychonomic Society*, 27(6):548–550, 1989. [33](#)
1039. H. Osman, M. Leuenberger, M. Lungu, and O. Nierstrasz. Tracking null checks in open-source Java systems. In *IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering*, SANER’16, pages 304–313, Mar. 2016. [169](#)
1040. L. M. Ottenstein, V. B. Schneider, and M. H. Halstead. Predicting the number of bugs expected in a program module. Technical Report CSD-TR 205, Purdue University, Oct. 1976. [164](#)
1041. M. A. Oumaziz, A. Charpentier, J.-R. Falleri, and X. Blanc. Documentation reuse: Hot or not? An empirical study. In *16th International Conference on Software Reuse*, ICSR 2017, pages 12–27, May 2017. [139](#)
1042. G. L. Ourada. Software cost estimating models: A calibration, validation, and comparison. Thesis (m.s.), Air Force Institute of Technology, Air University, USA, Dec. 1991. [5](#), [106](#)
1043. S. Owsowitz and A. Sweetland. Factors affecting coding errors. Research Memorandum RM-4346-PR, The RAND Corporation, Apr. 1965. [41](#)
1044. S. C. Özbek. *Introducing Innovations into Open Source Projects*. PhD thesis, Freie Universität Berlin, Aug. 2010. [114](#)
1045. A. Ozment and S. E. Schechter. Milk or wine: Does software security improve with age? In *USENIX Security Symposium (2006)*, pages 93–104, July-Aug. 2006. [135](#)
1046. P. Padfield. *Battleship*. Thistle Publishing, 2015. [2](#)
1047. R. Paleari, L. Martignoni, G. F. Roglia, and D. Bruschi. A fistful of red-pills: How to automatically generate procedures to detect CPU emulators. In *Proceedings of the 3rd USENIX conference on Offensive technologies*, WOOT’09, pages 2–2, Aug. 2009. [124](#)

1048. N. Palix, J. Lawall, and G. Muller. Tracking code patterns over multiple software versions with Herodotus. In *Proceedings of the 9th International Conference on Aspect-Oriented Software Development*, AOSD'10, pages 169–180, Mar. 2010. [129](#)
1049. N. Palix, S. Saha, G. Thomas, C. Calvès, J. Lawall, and G. Muller. Faults in Linux: Ten years later. Technical Report RR-7357, Institut National de Recherche en Informatique et en Automatique, Aug. 2010. [129](#)
1050. J. Pallister, S. Hollis, and J. Bennett. Identifying compiler options to minimise energy consumption for embedded platforms. In *eprint arXiv:cs.PF/1303.6485*, Aug. 2013. [327](#)
1051. S. E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, 1999. [33](#)
1052. H.-Y. Pan, A. Chao, and W. Foissner. A nonparametric lower bound for the number of species shared by multiple communities. *Journal of Agricultural, Biological, and Environmental Statistics*, 14(4):452–468, Dec. 2009. [99](#)
1053. K. Pan. *Using Evolution Patterns to Find Duplicated Bugs*. PhD thesis, Department of Computer Science, University of California at Santa Cruz, Oct. 2006. [139](#)
1054. A. Parkhomenko, A. Redkina, and O. Maslivets. Estimating hedonic price indexes for personal computers in Russia. MPRA Paper No. 5019, Higher School of Economics, Jan. 2007. [62](#)
1055. J. M. Parkman. Temporal aspects of simple multiplication and comparison. *Journal of Experimental Psychology*, 95(2):437–444, 1972. [41](#)
1056. J. M. Parkman and G. J. Groen. Temporal aspects of simple addition and comparison. *Journal of Experimental Psychology*, 89(2):335–342, 1971. [41](#)
1057. C. Parnin, C. Bird, and E. Murphy-Hill. Adoption and use of Java generics. *Empirical Software Engineering*, 18(6):1047–1089, Dec. 2013. [167](#)
1058. F. N. Parr. An alternative to the Rayleigh curve model for software development effort. *IEEE Transactions on Software Engineering*, SE-6(3):291–296, May 1980. [107](#)
1059. H. E. Pashler. *The Psychology of Attention*. The MIT Press, 1999. [43](#)
1060. L. Passos, J. Guo, L. Teixeira, K. Czarnecki, A. Wąsowski, and P. Borba. Coevolution of variability models and related artefacts: A case study of the Linux kernel. In *Proceedings of the 17th International Software Product Line Conference*, SPLC'13, pages 91–100, Apr. 2013. [84](#)
1061. A. Patel. Auditors' belief revision: Recency effects of contrary and supporting audit evidence and source reliability. Technical Report 2001-1, University of South Pacific, Dept. of AFM/SSE, June 2001. [27](#)
1062. M. R. Patterson. *Antitrust Law in the New Economy :Google, Yelp, LIBOR, and the Control of Information*. Harvard University Press, 2017. [78, 98](#)
1063. F. M. Paulus, L. Rademacher, T. A. J. Schäfer, L. Müller-Pinzler, and S. Krach. Journal impact factor shapes scientists' reward signal in the prospect of publication. *PLoS ONE*, 10(11):e0142537, Nov. 2015. [7](#)
1064. A. Pavese and C. Umiltà. Symbolic distance between numerosity and identity modulates stroop-like interference. *Journal of Experimental Psychology: Human Perception and Performance*, 24(5):1535–1545, 1998. [20](#)
1065. J. W. Payne, J. R. Bettman, and E. J. Bettman. *The Adaptive Decision Maker*. Cambridge University Press, 1993. [45](#)
1066. G. Paz-y-Miño C, A. B. Bond, A. C. Kamil, and R. P. Balda. Pinyon jays use transitive inference to predict social dominance. *Nature*, 430:778–781, Aug. 2004. [38](#)
1067. R. D. Pea. Language-independent conceptual "bugs" in novice programming. *Journal of Educational Computing Research*, 2(1):25–36, Feb. 1986. [150](#)
1068. J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000. [39](#)
1069. R. K. Pearson. The problem of disguised missing data. *ACM SIGKDD Explorations Newsletter*, 8(1):83–92, June 2006. [367](#)
1070. Y. Peers. *Econometric Advances in Diffusion Models*. PhD thesis, Erasmus University Rotterdam, Dec. 2011. [62](#)
1071. E. Pek, V. Klebanov, and R. Lämmel. Re-engineering software corpora for simplified adoption. ???, July 2017. [337](#)
1072. D. G. Pelli, C. W. Burns, B. Farell, and D. C. Moore. Identifying letters. *Vision Research*, 46(28):4646–4674, 2006. [34](#)
1073. E. Peltonen, E. Lagerspetz, P. Nurmi, and S. Tarkoma. Energy modeling of system settings: A crowdsourced approach. In *IEEE International Conference on Pervasive Computing and Communications*, PerCom, pages 37–45, Mar. 2015. [339](#)
1074. N. Pennington. Comprehension strategies in programming. In G. Olson, S. Shepard, and E. Soloway, editors, *Empirical Studies of Programmers: Second Workshop*, chapter 7, pages 100–113. Ablex Publishing, 1987. [154](#)
1075. N. Pennington. Stimulus structures and mental representations in expert comprehension of computer programs. *Cognitive Psychology*, 19(3):295–341, July 1987. [24, 168](#)
1076. C. Perez. *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*. Edward Elgar Publishing, 2003. [2, 58](#)
1077. D. E. Perry and W. M. Evangelist. An empirical study of software interface faults – An update. In *Proceedings of the Twentieth Annual Hawaii International Conference on Systems Sciences*, Vol II, pages 113–126, Jan. 1987. [6, 126](#)
1078. D. E. Perry, N. A. Staudenmayer, and L. G. Votta, Jr. Understanding and improving time usage in software development. In A. Fuggetta and A. L. Wolf, editors, *Trends in Software Process*, chapter 5, pages 111–135. John Wiley & Sons, Mar. 1995. [318](#)
1079. D. E. Perry and C. S. Stieg. Software faults in evolving a large, real-time system: a case study. In *Proceedings of the 1993 European Software Engineering Conference*, pages 48–67, Sept. 1993. [126](#)
1080. R. Perugupalli. Empirical assessment of architecture-based reliability of open-source software. Thesis (m.s.), Department of Computer Science and Electrical Engineering, West Virginia University, May 2004. [212](#)
1081. H. Petroski. *Design Paradigms: Case Histories of Error and Judgment in Engineering*. Cambridge University Press, 1994. [123](#)
1082. C. Peukert. Switching costs and information technology: The case of IT outsourcing. ???, ???(???)?:???, Aug. 2010. [120](#)
1083. A. Pewsey, M. Neuhauser, and G. D. Ruxton. *Circular Statistics in R*. Oxford University Press, 2013. [304, 305, 306](#)
1084. P. M. Pexman and M. J. Yap. Individual differences in semantic processing: Insights from the Calgary semantic decision project. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 44(7):1091–1112, Feb. 2018. [161](#)
1085. S. A. Phatak, A. Lovitt, and J. B. Allen. Consonant confusions in white noise. *Journal of the Acoustic Society of America*, 124(2):1220–1233, Aug. 2008. [162](#)
1086. C. Phillips. *Order and Structure*. PhD thesis, M.I.T., Aug. 1996. [22](#)
1087. M. Phister, Jr. *Data Processing Technology and Economics*. Santa Monica Publishing Company and Digital Press, second edition, 1979. [5, 9, 71, 77](#)
1088. S. T. Piantadosi. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review*, 21(5):1112–1130, Oct. 2014. [170](#)
1089. S. T. Piantadosi. A rational analysis of the approximate number system. *Psychonomic Bulletin & Review*, 23(3):877–886, June 2016. [40](#)
1090. R. Pieters and L. Warlop. Visual attention during brand choice: The impact of time pressure and task motivation. *International Journal of Research in Marketing*, 16:1–16, 1999. [33](#)
1091. D. J. Pigott and B. M. Axtens. Online historical encyclopedia of programming languages. <http://hopl.info/>, 2015. [89](#)
1092. J. Pipitone. Software quality in climate modelling. Thesis (m.s.), Department of Computer Science, University of Toronto, 2010. [131](#)
1093. A. M. Pires and C. ao Amado. Interval estimators for a binomial proportion: Comparison of twenty methods. *REVSTAT-Statistical Journal*, 6(2):165–197, June 2008. [230](#)
1094. D. J. Pittenger. Measuring the MBTI ... And coming up short. *Journal of Career Planning and Employment*, 54(1):48–52, Nov. 1993. [43](#)
1095. A. Pluchino, A. Rapisarda, and C. Garofalo. The Peter principle revisited: A computational study. In *eprint arXiv:physics.soc-ph/0907.0455v3*, Oct. 2009. [82](#)
1096. T. Plum. *Reliable data structures in C*. Plum Hall, 1985. [152](#)

1097. T. Plum. *C Programming guidelines*. Plum Hall, 1989. 152
1098. I. P. L. Png. On the reliability of software piracy statistics. *Electronic Commerce Research and Applications*, 9(5):365–373, Sept.-Oct. 2010. 63
1099. C. Poivey, J. L. Barth, K. A. LaBel, G. Gee, and H. Safren. In-flight observations of long-term single-event effect (SEE) performance on Orbview-2 solid state recorders (SSR). In *2003 IEEE Radiation Effects Data Workshop*, page ???, July 2003. 189
1100. R. Pollack. How to believe a machine-checked proof. In G. Sambin and J. M. Smith, editors, *Twenty Five Years of Constructive Type Theory*, chapter 11, pages 205–220. Oxford University Press, Oct. 1998. 124
1101. A. Pollatsek, E. D. Reichle, and K. Rayner. Tests of the E-Z reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, 52(1):1–56, Feb. 2006. 34
1102. D. Pope and U. Simonsohn. Round numbers as goals: Evidence from baseball, SAT takers, and the lab. *Psychological Science*, 22(1):71–79, Jan. 2011. 137
1103. K. R. Popper. *Conjectures and Refutations*. Routledge, 1969. 52
1104. A. Porter, H. Siy, A. Mockus, and L. Votta. Understanding the sources of variation in software inspections. *ACM Transactions on Software Engineering Methodology*, 7(1):41–79, Jan. 1998. 145, 259, 319
1105. M. E. Porter. The five competitive forces that shape strategy. *Harvard Business Review*, 86(1):78–93, Jan. 2008. 54, 58
1106. A. S. Posamentier and I. Lehmann. *Magnificent mistakes in mathematics*. Prometheus books, 2013. 124
1107. A. Potanin, M. Damitio, and J. Noble. Are your incoming aliases really necessary? Counting the cost of object ownership. In *Proceedings of the 2013 International Conference on Software*, ICSE’13, pages 742–751, May 2013. 325
1108. E. M. Pothos and N. Chater. Rational categories. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pages 848–853, 1998. 27, 311
1109. M. C. Potter, A. Moryadas, I. Abrams, and A. Noel. Word perception and misperception in context. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 19(1):3–22, 1993. 162
1110. J. Potts, J. Hartley, L. Montgomery, C. Neylon, and E. Rennie. A journal is a club: A new economic model for scholarly publishing. SSRN Working Paper n. 2763975, Apr. 2016. 8
1111. A. L. Powell. *Right on Time: Measuring, Modelling and Managing Time-Constrained Software Development*. PhD thesis, Department of Computer Science, University of York, Aug. 2001. 103, 293
1112. D. A. Powner and K. A. Rhodes. Business systems modernization: IRS needs to complete recent efforts to develop policies and procedures to guide requirements development and management. Technical Report GAO-06-310, United States Government Accountability Office, Mar. 2006. 118
1113. M. Pradel. *Program Analyses for Automatic and Precise Error Detection*. PhD thesis, ETH Zurich, 2012. 133
1114. L. Prechelt. The 28:1 Grant/Sackman legend is misleading, or: How large is interpersonal variation really? Technical Report iratr-1999-18, Universität Karlsruhe, 1999. 50, 74, 184
1115. L. Prechelt. Plat\_Forms 2007: The web development platform comparison-evaluation and results. Technical Report B-07-10, Institut für Informatik, Freie Universität Berlin, June 2007. 113, 173
1116. L. Prechelt, D. Graziotin, and D. M. Fernández. On the status and future of peer review in software engineering. In *eprint arXiv:cs.SE/1706.07196*, June 2017. 7
1117. L. Prechelt and W. F. Tichy. A controlled experiment measuring the effect of procedure argument type checking on programmer productivity. *IEEE Transactions on Software Engineering*, 24(4):302–312, Apr. 1998. 153
1118. L. Prechelt, F. Zieris, and H. Schmeisky. Difficulty factors of obtaining access for empirical studies in industry. In *Proceedings of the Third International Workshop on Conducting Empirical Studies in Industry*, CESI’15, pages 19–25, May 2015. 5
1119. L. S. Premo and S. L. Kuhn. Modeling effects of local population extinctions on cultural change and diversity in the paleolithic. *PLoS ONE*, 5(12):e15582, Dec. 2010. 71, 192
1120. C. C. Presson and D. R. Montello. Updating after rotational and translational body movements: coordinate structure of perspective space. *Perception*, 23:1447–1455, 1994. 17
1121. D. Pritchard. Frequency distribution of error messages. In *Proceedings of the 6th Workshop on Evaluation and Usability of Programming Languages and Tools*, PLATEAU 2015, pages 1–8, Oct. 2015. 139
1122. V. Propp. *Morphology of the Folktale*. University of Texas Press, second edition, 1968. English translation by Laurence Scott. 168
1123. J. Prümper, D. Zapf, F. C. Brodbeck, and M. Frese. Some surprising differences between novice and expert errors in computerized office work. *Behaviour & Information Technology*, 11(6):319–328, 1992. 123
1124. R. Purushothaman and D. E. Perry. Toward understanding the rhetoric of small source code changes. *IEEE Transactions on Software Engineering*, 31(6):511–526, June 2005. 140, 141
1125. L. H. Putnam. A general empirical solution to the macro software sizing and estimating problem. *IEEE Transactions on Software Engineering*, SE-4(4):345–361, July 1978. 107
1126. L. H. Putnam and W. Myers. *Measures for Excellence: Reliable software on time, within budget*. Prentice-Hall, Inc, 1992. 195
1127. PwC. Converging forces are building that could re-shape the entire industry. Global 100 software leaders, PwC Technology Institute, May 2013. 64
1128. PwC. The growing importance of apps and services. Global 100 software leaders, PwC Technology Institute, Mar. 2014. 64
1129. PwC. Digital intelligence conquers the world below and the cloud above. Global 100 software leaders, PwC Technology Institute, 2016. 64
1130. PwC. IPO review full-year and Q4 2015. Global technology, PwC Technology Institute, Feb. 2016. 80
1131. Z. Pylyshyn. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences*, 22(3):341–423, 1999. 32
1132. X. Qu. *Configuration aware prioritization for regression testing*. PhD thesis, The Graduate College at the University of Nebraska, Apr. 2010. 147
1133. S. Qualline. *C Elements of Style*. M&T Books, 1992. 152
1134. R. Queiroz, L. Passos, M. T. Valente, C. Hunsen, S. Apel, and K. Czarnecki. The shape of feature code: an analysis of twenty C-preprocessor-based systems. *Journal on Software and Systems Modeling*, 16(1):77–96, Feb. 2017. 276, 277
1135. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. ISBN 3-900051-07-0. 2, 349, 350
1136. R Core Team. R language definition. Technical Report 3.3.1, R Foundation for Statistical Computing, June 2016. 349
1137. H. Rabinowitz and C. Schaap. *Portable C*. Prentice-Hall, Inc, 1990. 152
1138. J. W. Radatz. Analysis of IV & V data. Technical Report RADC-TR-81-145, Rome Air Development Center, Griffiss Air Force Base, June 1981. 138
1139. G. Radden and R. Dirven. *Cognitive English Grammar*. John Benjamins Publishing Company, 2007. 149
1140. D. Raffo, J. Settle, and W. Harrison. Investigating financial measures for planning software IV&V. Technical Report TR-99-05, Portland State University, 1999. 56
1141. C. Ragkhitwetsagul, J. Krinke, and D. Clark. A comparison of code similarity analysers. *Empirical Software Engineering*, 23(4):2464–2519, Aug. 2018. 165
1142. F. Rahman, C. Bird, and P. Devanbu. Clones: What is that smell? In *Proceedings of the 7th International Workshop on Mining Software Repositories*, MSR’10, pages 72–81, May 2010. 6
1143. J. Ranade and A. Nash. *The Elements of C Programming Style*. McGraw-Hill, Inc, 1992. 152
1144. B. Ray. *Analysis of Cross-System Porting and Porting Errors in Software Projects*. PhD thesis, University of Texas at Austin, Aug. 2013. 86, 87, 129
1145. K. Rayner. Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8):1457–1506, 2009. 34

1146. K. Rayner. Eye movements in reading: Models and data. *Journal of Eye Movement Research*, 2(5):1–10, Apr. 2009. 33
1147. N. M. Razali and Y. B. Wah. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011. 227
1148. J. Reason. *Human Error*. Cambridge University Press, 1990. 41, 123, 136
1149. A. S. Reber and S. M. Kassin. On the relationship between implicit and explicit modes in the learning of a complex rule structure. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):492–502, 1980. 20, 24
1150. B. Regnell, M. Höst, J. N. och Dag, P. Beremark, and T. Hjelm. An industrial case study on distributed prioritisation in market-driven requirements engineering for packaged software. *Requirements Engineering*, 6(1):51–62, Apr. 2001. 47, 118
1151. E. D. Reichle, T. Warren, and K. McConnell. Using E-Z reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, 16(1):1–21, Feb. 2009. 34
1152. G. A. Reis III. *Software Modulated Fault Tolerance*. PhD thesis, Department of Electrical Engineering, Princeton University, June 2008. 142
1153. G. Remillard. Implicit learning of second-, third-, and fourth-order adjacent and nonadjacent sequential dependencies. *The Quarterly Journal of Experimental Psychology*, 61(3):400–424, Apr. 2008. 20
1154. R. W. Remington, H. W. H. Yuen, and H. Pashler. With practice, keyboard shortcuts become faster than menu selection: A crossover interaction. *Journal of Experimental Psychology: Applied*, 22(1):95–106, 2016. 41
1155. Research Councils UK. RCUK policy on open access and supporting guidance. Technical Report ???, RCUK, Apr. 2013. 7
1156. G. Richards, C. Hammer, B. Burg, and J. Vitek. The eval that men do A large-scale study of the use of eval in JavaScript applications. In *Proceedings of the 25th European conference on Object-oriented programming*, ECOOP’11, pages 52–78, July 2011. 165
1157. G. Richards, S. Lebresne, B. Burg, and J. Vitek. An analysis of the dynamic behavior of JavaScript programs. In *ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI’10, pages 1–12, June 2010. 165
1158. R. Richardson. 2008 CSI computer crime & security survey. Technical report, Computer Security Institute, Aug. 2008. 129
1159. D. F. Rico. Short history of software methods. <http://davidfrico.com/rico04e.pdf>, July 2004. 114
1160. R. Riesen, K. Ferreira, J. Stearley, R. Oldfield, J. H. Laros III, K. Pedretti, and R. Brightwell. Redundant computing for exascale systems. Technical Report SAND2010-8709, Sandia National Laboratories, Dec. 2010. 142
1161. M. Rinard, C. Cadar, and H. H. Nguyen. Exploring the acceptability envelope. In *Companion to the 20th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications*, OOPSLA’05, pages 21–30, Oct. 2005. 123
1162. 7digital, ltd??? website, July 2012. <http://www.7digital.com>. 119, 120, 208, 288, 289, 290, 366
1163. R. Robbes, D. Röthlisberger, and É. Tanter. Object-oriented software extensions in practice. *Empirical Software Engineering*, 20(3):745–782, June 2015. 171
1164. M. J. Roberts, D. J. Gilmore, and D. J. Wood. Individual differences and strategy selection in reasoning. *British Journal of Psychology*, 88:473–492, 1997. 36
1165. S. Roberts and J. Winters. Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits. *PLoS ONE*, 8(8):e70902, Aug. 2013. 229
1166. D. E. Robinson. Fashions in shaving and trimming of the beard: The men of the Illustrated London News, 1842–1972. *American Journal of Sociology*, 81(5):1133–1141, Mar. 1976. 6
1167. G. Robles and J. M. González-Barahona. A comprehensive study of software forks: Dates, reasons and outcomes. In *The 8th International Conference on Open Source Systems*, OSS 2012, pages 1–14, Sept. 2012. 86
1168. G. Robles, I. Herraiz, D. M. Germán, and D. Izquierdo-Cortázar. Modification and developer metrics at the function level: Metrics for the study of the evolution of a software project. In *3rd International Workshop on Emerging Trends in Software Metrics*, WETSoM, pages 49–55, June 2012. 94, 95
1169. G. Robles, L. A. Reina, A. Serebrenik, B. Vasilescu, and J. M. González-Barahona. FLOSS 2013: A survey dataset about free software contributors: Challenges for curating, sharing, and combining. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, MSR’14, pages 396–399, May 2014. 180, 346
1170. E. Rodrigues Jr. and R. Terra. How do developers use dynamic features? The case of Ruby. *Computer Languages, Systems & Structures*, 53(?):73–89, Sept. 2018. 169
1171. W. H. Roetzheim. When the software becomes a nightmare: Dealing with failed projects. *Business Law Today*, 13(6):42–48, July-Aug. 2004. 115
1172. R. D. Rogers and S. Monsell. Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, 124(2):207–231, 1995. 44
1173. E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, July 1976. 28
1174. M. Rosenfelder. *The Language Construction Kit*. Yonagu Books, 2010. 89
1175. A. Ross. *No-Collar: The Humane Workplace and its Hidden Costs*. Temple University Press, 2003. 80
1176. B. H. Ross and G. L. Murphy. Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 38(4):495–552, June 1999. 313
1177. L. Ross, M. R. Lepper, and M. Hubbard. Perseverance in self-perception and social perception: Biased attributional processes in the debelieving paradigm. *Journal of Personality and Social Psychology*, 32(5):880–892, 1975. 26
1178. J. Rost and R. L. Glass. *The Dark Side of Software Engineering: Evil on Computing Projects*. John Wiley & Sons, Inc, 2011. 102, 117
1179. V. Rothberg, N. Dintzner, A. Ziegler, and D. Lohmann. Feature models in Linux—from symbols to semantics. In *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems*, VaMoS’16, pages 65–72, Jan. 2016. 84, 85
1180. B. F. Roukema. A first-digit anomaly in the 2009 Iranian presidential election. In *eprint arXiv:stat.AP/0906.2789*, June 2013. 372
1181. M. M. Roy, N. J. S. Christenfeld, and C. R. M. McKenzie. Underestimating the duration of future events: Memory incorrectly used or memory bias? *Psychological Bulletin*, 131(5):738–756, 2005. 48
1182. W. W. Royce. Managing the development of large software systems. In *Proceedings IEEE WESON*, pages 1–9, Aug. 1970. 113
1183. P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, Jan. 2006. 260
1184. D. C. Rubin and A. E. Wenzel. One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, 103(4):734–760, 1996. 23
1185. C. Rubio-González and B. Libit. Expect the unexpected: Error code mismatches between documentation and the real world. In *Proceedings of the 9th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software tools and engineering*, PASTE’10, pages 73–80, June 2010. 141
1186. C. Rubio-González, C. Nguyen, H. D. Nguyen, J. Demmel, W. Kahan, K. Sen, D. H. Bailey, C. Iancu, and D. Hough. Precimoniou: Tuning assistant for floating-point precision. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC’13, Nov. 2013. 125
1187. A. L. Russell. ‘rough consensus and running code’ and the internet-OSI standards war. *IEEE Annals of the History of Computing*, 28(3):48–61, July-Sept. 2006. 113
1188. R. Sabherwal, A. Jeyaraj, and C. Chowdhury. Information systems success: Individual and organizational determinants. *Management Science*, 52(12):1849–1864, Dec. 2006. 235
1189. R. Saborido, V. Arnaoudova, G. Beltrame, F. Khomh, and G. Antoniol. On the impact of sampling frequency on software energy measurements. *PeerJ PrePrints*, 3:e1219, July 2015. 339
1190. H. Sackman, W. J. Erikson, and E. E. Grant. Exploratory experimental studies comparing online and offline programming performance. *Communications of the ACM*, 11(1):3–11, Jan. 1968. 74

1191. M. Sadat, A. B. Bener, and A. V. Miranskyy. Rediscovery datasets: Connecting duplicate reports. In *eprint arXiv:cs/1703.06337v1*, Mar. 2017. [126](#), [135](#)
1192. M. Sadinle. On the performance of dual system estimators of population size: A simulation study. Documentos de CERAC No. 13, Centro de Recursos para el Análisis de Conflictos, Bogotá, Columbia, Dec. 2008. [98](#)
1193. S. K. Sahoo, J. Criswell, and V. Adve. An empirical study of reported bugs in server software with implications for automated bug diagnosis. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering, ICSE'10*, pages 485–494, May 2010. [127](#)
1194. J. Sajaniemi and R. N. Prieto. Roles of variables in experts' programming knowledge. In *17th Workshop of the Psychology of Programming Interest Group*, pages 145–159, June 2005. [170](#)
1195. P. H. Salus. *A Quarter Century of UNIX*. Addison-Wesley, 1994. [91](#)
1196. P. H. Salus. Duelling UNIXes and the UNIX wars. *;login:*, 40(2):66–68, Apr. 2015. [91](#)
1197. Like everyone else, Twitter hides from U.S. taxes in Ireland. website: accessed: 5-Feb-17, Oct. 2013. <http://valleywag.gawker.com/like-everyone-else-twitter-hides-from-u-s-taxes-in-ir-1447085830>. [60](#)
1198. A. Sampson, W. Dietl, E. Fortuna, and D. Gnanapragasam. EnerJ: Approximate data types for safe and general low-power computation. In *Proceedings of the 32nd ACM SIGPLAN conference on Programming language design and implementation, PLDI'11*, pages 164–174, June 2011. [338](#)
1199. D. M. Sanbonmatsu, S. S. Posavac, A. A. Behrends, S. M. Moore, and B. N. Uchino. Why a confirmation strategy dominates psychological science. *PLoS ONE*, page e0138197, Sept. 2015. [52](#)
1200. D. Sarkar. *Lattice Multivariate Data Visualization with R*. Springer Science+Business Media, 2008. [189](#)
1201. M. Savić, M. Ivanović, Z. Budimac, and M. Radovanović. Do students' programming skills depend on programming language? In *International Conference of Numerical Analysis and Applied Mathematics, ICNAAM 2015*, page ???, Apr. 2015. [91](#)
1202. S. R. Schach, T. O. S. Adeshian, D. Balasubramanian, G. Madl, E. P. Osses, S. Singh, K. Suwanmongkol, M. Xie, and D. G. Feitelson. Common coupling and pointer variables, with application to a Linux case study. *Software Quality Journal*, 15(1):99–113, Mar. 2006. [145](#)
1203. S. R. Schach, B. Jin, L. Yu, G. Z. Heller, and J. Offutt. Determining the distribution of maintenance categories: Survey versus measurement. *Empirical Software Engineering*, 8(4):351–363, Dec. 2003. [85](#), [282](#)
1204. J. Schad, J. Dittrich, and J.-A. Quiané-Ruiz. Runtime measurements in the cloud: Observing, analyzing, and reducing variance. In *Proceedings of the VLDB Endowment*, pages 460–471, Sept. 2010. [341](#), [345](#)
1205. K. W. Schaie. *Developmental Influences on Adult Intelligence: The Seattle Longitudinal Study*. Oxford University Press, second edition, 2013. [49](#)
1206. R. R. Schaller. *Technological Innovation in the Semiconductor Industry: A Case Study of the International Technology Roadmap for Semiconductors (ITRS)*. PhD thesis, George Mason University, 2004. [68](#)
1207. E. Schneider, M. Maruyama, S. Dehaene, and M. Sigman. Eye gaze reveals a fast, parallel extraction of the syntax of arithmetic formulas. *Cognition*, 125(3):475–490, Dec. 2012. [34](#)
1208. J. Schock, M. J. Cortese, M. M. Khanna, and S. Toppi. Age of acquisition estimates for 3,000 disyllabic words. *Behavior and Research Methods*, 44(4):971–977, Dec. 2012. [161](#)
1209. A. Scholey and L. Owen. Effects of chocolate on cognitive function and mood: a systematic review. *Nutrition Reviews*, 71(10):665–681, Apr. 2013. [50](#)
1210. R. Schöne, D. Hackenberg, and D. Molka. Memory performance at reduced CPU clock speeds: An analysis of current x86\_64 processors. In *Proceedings of the 2012 USENIX conference on Power-Aware Computing and Systems, HotPower'12*, Oct. 2012. [185](#), [341](#)
1211. M. Schonlau, W. DuMouchel, W.-H. Ju, A. F. Karr, M. Theus, and Y. Vardi. Computer intrusion: Detecting masquerades. *Statistical Science*, 16(1):58–74, 2001. [75](#)
1212. L. J. Schooler and R. Hertwig. How forgetting aids heuristic inference. *Psychological Review*, 112(3):610–628, 2005. [23](#)
1213. E. R. Schotter, B. Angele, and K. Rayner. Parafoveal processing in reading. *Attention, Perception & Psychophysics*, 74(1):5–35, Jan. 2012. [34](#)
1214. J.-P. Schraepler and G. G. Wagner. Identification of faked interviews in surveys by means of Benford's law?: An analysis by means of genuine fakes in the raw data of SOEP. Technical report, Technische Universität Berlin, Aug. 2004. [372](#)
1215. M.-A. Schulz, B. Schmalbach, P. Brugger, and K. Witt. Analysing humanly generated random number sequences: A pattern-based approach. *PLoS ONE*, 7(7):e41531, July 2012. [373](#)
1216. P. Schuurman, E. Berghout, and P. Powell. Benefits are from Venus, costs are from Mars. CITER WP/010/PSEBPP, University of Groningen Centre for IT Economics Research, June 2008. [101](#)
1217. C. F. Scott, P. Cole, R. B. Hesse, and P. R. Malone. UNITED STATES OF AMERICA, et al., v. ORACLE CORPORATION. Plaintiff's post-trial brief CASE NO. C 04-0807 VRW, UNITED STATES DISTRICT COURT NORTHERN DISTRICT OF CALIFORNIA SAN FRANCISCO DIVISION, July 2004. [98](#)
1218. P. D. Scott and M. Fasli. Benford's law: An empirical investigation and a novel explanation. CSM Technical Report 349, Department of Computer Science, University of Essex, Aug. 2001. [372](#)
1219. S. Scribner. Modes of thinking and ways of speaking: culture and logic reconsidered. In P. N. Johnson-Laird and P. C. Wason, editors, *Thinking: Readings in Cognitive Science*, chapter 29, pages 483–500. Cambridge University Press, 1977. [35](#)
1220. R. C. Seacord. *The CERT C Secure Coding Standard*. Addison-Wesley, 2009. [152](#)
1221. R. C. Seamans, Jr. *Aiming at Targets: The Autobiography of Robert C. Seamans, Jr.* NASA History Office, 1996. [102](#)
1222. The world's largest hedge fund is a fraud. SEC MADOFF EXHIBITS-04451, Nov. 2005. November 7, 2005 Submission to the SEC, Madoff Investment Securities, LLC. [373](#)
1223. P. Sehgal, V. Tarasov, and E. Zadok. Evaluating performance and energy in file system server workloads. In *Proceedings of the 8th USENIX conference on File and storage technologies, FAST'10*, Feb. 2010. [343](#)
1224. J. Selby and K. Mayer. Startup firm acquisitions as a human resource strategy for innovation: The acquire phenomenon. *Academy of Management Proceedings*, 1:17109–17109, Nov. 2013. [80](#)
1225. R. W. Selby, Jr., V. R. Basili, and F. T. Baker. CLEANROOM software development: An empirical evaluation. Technical Report TR-1415, Department of Computer Science, University of Maryland, Feb. 1985. [109](#)
1226. L. L. Selwyn. *Economies of Scale in Computer Use: Initial Tests and Implications for the Computer Utility*. PhD thesis, Alfred P. Sloan School of Management, June 1969. [77](#)
1227. J. A. Sexton. Detecting errors in software using a parameter checker: An analysis. Thesis (m.s.), Rochester Institute of Technology, Apr. 1989. [139](#)
1228. N. Shadbolt. Shadbolt review of computer sciences degree accreditation and graduate employability. Technical Report IND/16/5, Department for Business, Innovation & Skills, UK, Apr. 2016. [81](#)
1229. T. M. Shaft and I. Vessey. The relevance of application domain knowledge: Characterizing the computer program comprehension process. *Journal of Management Information Systems*, 15(1):51–78, 1998. [122](#), [154](#)
1230. J. Shallit. Randomized algorithms in "primitive" cultures or what is the oracle complexity of a dead chicken? *ACM SIGACT News*, 23(4):77–80, 1992. [154](#)
1231. C. Shaoul, R. H. Baayen, and C. F. Westbury. N-gram probability effects in a cloze task. *The Mental Lexicon*, 9(3):437–472, 2014. [162](#)
1232. C. Shapiro and H. R. Varian. The art of standards wars. *California Management Review*, 41(2):8–32, Jan. 1999. [80](#)
1233. W. F. Sharpe. *The Economics of Computers*. Columbia University Press, 1969. [77](#)
1234. O. Shatnawi. Measuring commercial software operational reliability: an interdisciplinary modelling approach. *Eksplotacja i Niezawodność – Maintenance and Reliability*, 16(4):585–594, 2014. [131](#)

1235. D. E. Shaw, R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC'09, page 39, Nov. 2009. 77
1236. B. R. Shear and B. D. Zumbo. False positives in multiple regression: Unanticipated consequences of measurement error in the predictor variables. *Educational and Psychological Measurement*, 73(5):733–756, Oct. 2013. 244
1237. D. Shefer. Pricing for software product managers. ???, 2005. 61
1238. B. A. Sheil. The psychological study of programming. *ACM Computing Surveys*, 13(1):101–120, Mar. 1981. 6
1239. T.-J. Shen, A. Chao, and C.-F. Lin. Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3):798–804, Mar. 2003. 99
1240. V. Y. Shen, S. D. Conte, and H. E. Dunsmore. Software science revisited: A critical analysis of the theory and its empirical support. Technical Report CSD-TR-375, Purdue University, Jan. 1981. 164
1241. A. Shenhav, S. Musslick, F. Lieder, W. Kool, T. L. Griffiths, J. D. Cohen, and M. M. Botvinick. Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience*, 40:99–124, July 2017. 42
1242. R. N. Shepard, C. I. Hovland, and H. M. Jenkins. Learning and memorization of classifications. *Psychological Monographs: General and Applied*, 75(15):1–39, 1961. 29
1243. R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, Feb. 1971. 18
1244. S. B. Sheppard and E. Kruesi. The effects of the symbology and spatial arrangement of software specifications in a coding task. Technical Report TR-81-388200-3, Information Systems Programs, General Electric, Feb. 1981. 138
1245. L. Shi, H. Zhong, T. Xie, and M. Li. An empirical study on evolution of API documentation. In *Proceedings of the 14th international conference on Fundamental approaches to software engineering*, FASE'11/ETAPS'11, pages 416–431, Apr. 2011. 88
1246. E. Shihab, A. Ihara, Y. Kamei, W. M. Ibrahim, M. Ohira, B. Adams, A. E. Hassan, and K. ichi Matsumoto. Predicting re-opened bugs: A case study on the Eclipse project. In *17th Working Conference on Reverse Engineering*, WCRE, pages 249–258, Oct. 2010. 310
1247. E. Shihab, Z. M. Jiang, W. M. Ibrahim, B. Adams, and A. E. Hassan. Understanding the impact of code and process metrics on post-release defects: A case study on the Eclipse project. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM'10, pages 1–4, Sept. 2010. 140
1248. M. Shimasaki, S. Fukaya, K. Ikeda, and T. Kiyono. An analysis of Pascal programs in compiler writing. *Software—Practice and Experience*, 10(2):149–157, Feb. 1980. 109
1249. T. C. Shrum. Calibration and validation of the checkpoint model to the air force electronic systems center software database. Thesis (m.s.), Graduate School of Logistics and Acquisition Management or the Air Force Institute of Technology, Air University, Sept. 1997. 5
1250. A. Shtub, N. Levin, and S. Globerson. Learning and forgetting industrial skills: An experimental model. *The International Journal of Human Factors in Manufacturing*, 3(3):293–305, July 1993. 26
1251. O. Shy. *How to Price: A Guide to Pricing Techniques and Yield Management*. Cambridge University Press, 2008. 61
1252. N. Sigmund, M. Rosenmüller, C. Kästner, P. G. Giarrusso, S. Apel, and S. S. Kolesnikov. Scalable prediction of non-functional properties in software product lines: Footprint and memory consumption. *Information and Software Technology*, 55(3):491–507, Mar. 2013. 168
1253. T. Simcoe. Standard setting committees: Consensus governance for shared technology platforms. *American Economic Review*, 102(1):305–336, Feb. 2013. 80
1254. T. Simcoe. Modularity and the evolution of the internet. In A. Goldfarb, S. M. Greenstein, and C. E. Tucker, editors, *Economic Analysis of the Digital Economy*, chapter 1, pages 21–47. University of Chicago Press, May 2015. 112, 113
1255. T. S. Simcoe and D. M. Wagquespack. Status, quality and attention: What's in a (missing) name? *Management Science*, 57(2):274–290, Sept. 2011. 72
1256. K. M. Simmons and D. Sutter. False alarms, tornado warnings, and tornado casualties. *Weather, Climate, and Society*, 1(1):38–53, Oct. 2009. 197
1257. H. A. Simon. *Models of Bounded Rationality: Behavioral Economics and Business Organization*. The MIT Press, 1982. 44
1258. H. A. Simon. Making management decisions: the role of intuition and emotion. *The Academy of Management Executive* (1987–1989), 1(1):57–64, Feb. 1987. 26
1259. I. Simonson. Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research*, 16:158–173, Sept. 1989. 47
1260. I. C. Simpson, P. Mousikou, J. M. Montoya, and S. Defior. A letter visual-similarity matrix for Latin-based alphabets. *Behavior and Research Methods*, 45(2):431–439, June 2013. 42
1261. J. Singer, M. Luján, and I. Watson. Meaningful type names as a basis for object lifetime prediction. In ???, OOPSLA'08, page ???, Apr. 2008. 167
1262. D. I. K. Sjøberg, B. Anda, E. Arisholm, T. Dybå, M. Jørgensen, A. Karahasanović, E. F. Koren, and M. Vokáč. Conducting realistic experiments in software engineering. In *Proceedings of the 2002 International Symposium on Empirical Software Engineering*, ISESE'02, pages 17–26, Oct. 2002. 320
1263. D. I. K. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanović, N.-K. Liborg, and A. C. Rekdal. A survey of controlled experiments in software engineering. Technical Report 2004-4, SIMULA Research Laboratory, 2004. 318
1264. I. Skoulis. Analysis of schema evolution for databases in open-source software. Thesis (m.s.), University of Ioannina, Greece, Sept. 2013. 96
1265. G. Slade. *Made to Break*. Harvard University Press, 2007. 4
1266. S. A. Slaughter, S. Ang, and W. F. Boh. Firm-specific human capital and compensation-organizational tenure profiles: An archival analysis of salary data for IT professionals. *Human Resource Management*, 46(3):373–394, 2007. 82
1267. S. Sloman, A. K. Barbey, and J. M. Hotaling. A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, 33(1):21–50, Jan. 2009. 39
1268. S. A. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22, 1996. 36
1269. S. A. Sloman. Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35(1):1–33, Feb. 1998. 28
1270. S. A. Sloman, M. C. Harrison, and B. C. Malt. Recent exposure affects artifact naming. *Memory & Cognition*, 30(5):687–695, 2002. 73
1271. S. A. Sloman and D. Lagnado. Causality in thought. *Annual Review of Psychology*, 66:223–247, 2015. 39
1272. S. A. Sloman and D. A. Lagnado. Do we "do"? *Cognitive Science*, 29(1):5–39, Jan.–Feb. 2005. 39
1273. P. Slovic. *The Perception of Risk*. Earthscan Publications Ltd, 2000. 127
1274. P. E. Smaldino and R. McElreath. The natural selection of bad science. *Royal Society Open Science*, 3:160384, Aug. 2016. 8
1275. G. K. Smith, A. A. Barbour, T. L. McNaugher, M. D. Rich, and W. L. Stanley. The use of prototypes in weapon system development. Report R-2345-AF, The RAND Corporation, Mar. 1981. 117
1276. F. Söhnchen and S. Albers. Pipeline management for the acquisition of industrial projects. *Industrial Marketing Management*, 39(8):1356–1364, Nov. 2010. 109
1277. M. Sojer, O. Alexy, S. Kleinknecht, and J. Henkel. Understanding the drivers of unethical programming behavior: The inappropriate reuse of internet-accessible code. *Journal of Management Information Systems*, 31(3):287–325, 2014. 111
1278. Solganick & Co. Software M&A update. <http://www.solganickco.com/wp-content/uploads/2017/02/Solganick-Software-Q4-2016-final.pdf>, Apr. 2016. 80
1279. G. S. Sommer. *Astronomical Odds A Policy Framework for the Cosmic Impact Hazard*. PhD thesis, Pardee RAND Graduate School, U.S.A., June 2004. 127

1280. J. Sonnemanns. Price clustering and natural resistance points in the Dutch stock market: A natural experiment. *European Economic Review*, 50(8):1937–1950, Nov. 2006. 137
1281. R. W. Soukoreff. *Quantifying Text Entry Performance*. PhD thesis, York University, Toronto, Canada, Apr. 2010. 41
1282. SPEC. Standard performance evaluation corporation. <http://spec.org>, July 2014. 181, 231, 264, 337
1283. SPEC. SPEC power\_ssj 2008. [http://spec.org/power\\_ssj2008](http://spec.org/power_ssj2008), June 2016. 271
1284. I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, Apr. 1991. 188, 189
1285. M. Spence. Job market signalling. *The Quarterly Journal of Economics*, 87(3):355–374, Aug. 1973. 80
1286. D. Sperber and D. Wilson. *Relevance: Communication and Cognition*. Blackwell Publishers, second edition, 1995. 36, 149
1287. D. Spinellis. *Code Reading: The Open Source Perspective*. Addison-Wesley, 2003. 152
1288. D. Spinellis, V. Karakoidas, and P. Louridas. Comparative language fuzz testing: Programming languages vs. fat fingers. In *Proceedings of the Workshop on Evaluation and Usability of Programming Languages and Tools*, PLATEAU, pages 25–34, Oct. 2012. 139
1289. J. Spolsky. Fog Creek professional ladder. <https://www.joelonsoftware.com/2009/02/13/fog-creek-professional-ladder>, Feb. 2009. 82
1290. J. Sprouse and D. Almeida. Assessing the reliability of textbook data in syntax: Adger’s core syntax. *Journal of Linguistics*, 48(3):609–652, Nov. 2012. 137
1291. D. Spuler. *C++ and C debugging, testing and reliability*. Prentice-Hall, Inc, 1994. 152
1292. L. R. Squire and A. J. O. Dede. Conscious and unconscious memory systems. *Perspectives in Biology*, 7(3):a021667, Mar. 2015. 19
1293. J. Srinivasan. *Lifetime Reliability Aware Microprocessor*. PhD thesis, University of Illinois at Urbana-Champaign, Oct. 2006. 141
1294. E. B. Staats. Millions in savings possible in converting programs from one computer to another. Technical Report FGMSD-77-34, Office of Management and Budget, National Bureau of Standards, Sept. 1977. 89
1295. The dirty little secret of software pricing. website, 2012. [http://www.rti.com/whitepapers/Dirty\\_Little\\_Secret.pdf](http://www.rti.com/whitepapers/Dirty_Little_Secret.pdf). 61
1296. Standish Group. The CHAOS report. Technical report, The Standish Group International, Inc, Aug. 1994. 104
1297. P. Stanley-Marbell, V. Estellers, and M. Rinard. Crayon: Saving power through shape and color approximation on next-generation displays. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys’16, page ???, Apr. 2016. 339
1298. K. E. Stanovich. *Who Is Rational? Studies of Individual Differences in Reasoning*. Lawrence Erlbaum Associates, 1999. 36
1299. K. E. Stanovich and R. F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23(5):645–726, Oct. 2000. 36
1300. M. Staples, R. Kolanski, G. Klein, C. Lewis, J. Andronick, T. Murray, R. Jeffery, and L. Bass. Formal specifications better than function points for code sizing. In *International Conference on Software Engineering*, ICSE 2013, pages 1257–1260, May 2013. 176
1301. J. Starek. A large-scale analysis of Java API usage. Thesis (m.s.), Institut für Informatik, Universität Koblenz-Landau, Mar. 2010. 260
1302. T. N. Starr, L. K. Picton, and J. W. Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549:409–413, Sept. 2017. 68
1303. M. Steele and J. Chaseling. Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions. *Communications in Statistics-Simulation and Computation*, 35(4):1067–1075, Apr. 2006. 206
1304. R. G. Steen, A. Casadevall, and F. C. Fang. Why has the number of scientific retractions increased? *PLoS ONE*, 8(7), Apr. 2013. 8
1305. J. Steffens. *Newgames: Strategic Competition in the PC revolution*. Pergamon Press, 1994. 78
1306. T. Stengos and E. Zacharias. Intertemporal pricing and price discrimination: A semiparametric hedonic analysis of the personal computer market. Discussion Paper 2002-11, Department of Economics, University of Cyprus, June 2002. 63
1307. K. Stenning and M. van Lambalgen. Semantics as a foundation for psychology: A case study of Wason’s selection task. *Journal of Logic, Language and Information*, 10(3):273–317, June 2001. 36
1308. K. Stenning and M. van Lambalgen. A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science*, 28(4):481–530, July-Aug. 2004. 36
1309. K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008. 36
1310. M. A. Stephens. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69(347):730–737, Sept. 1974. 204
1311. R. J. Sternberg and E. M. Weil. An aptitude-strategy interaction in linear syllogistic reasoning. Technical Report 15, Department of Psychology, Yale University, Apr. 1979. 36
1312. S. Sternberg. Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, 57(4):421–457, 1969. 22
1313. A. Stevens and P. Coupe. Distortions in judged spatial relations. *Cognitive Psychology*, 10(4):422–437, Oct. 1978. 28
1314. N. Stewart, N. Chater, and G. D. Brown. Decision by sampling. *Cognitive Psychology*, 53(1):1–26, Jan. 2006. 138, 197
1315. N. Stewart, C. Ungemach, A. J. L. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10(5):479–491, Sept. 2015. 48, 321
1316. G. Stikkel. Dynamic model for the system testing process. *Information and Software Technology*, 48(7):578–585, July 2006. 145
1317. V. Stodden, P. Guo, and Z. Ma. Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS ONE*, 8(6):e13636, June 2013. 7
1318. G. P. Stone, D. B. Levin, H. Hwang, M. Kim, and C. Mckay. JANET SKOLD and DAVID DOSSANTOS, on behalf of themselves and all others similarly situated, v. INTEL CORPORATION, HEWLETT PACKARD COMPANY and DOES 1-50, case no. 1-05-CV-039231, filing #g-64475. Opinion, Superior court of the state of California for the county of Santa Clara, 2014. 336
1319. J. Stone, M. Greenwald, C. Partridge, and J. Hughes. Performance of checksums and CRCs over real data. *IEEE/ACM Transactions on Networking*, 6(5):529–543, Oct. 1998. 126
1320. D. Straker. *C-Style standards and guidelines*. Prentice-Hall, Inc, 1992. 152
1321. S. Strand, I. J. Deary, and P. Smith. Sex differences in cognitive abilities test scores: A UK national picture. *British Journal of Educational Psychology*, 76(3):463–480, Apr. 2006. 16, 17
1322. H. . Strong. Mozilla foundation and subsidiary december 31, 2015 and 2014. Independent auditors’ report and consolidated financial statements, Hood & Strong LLC, Nov. 2016. 103
1323. R. Sudan, S. Ayers, P. Dongier, A. Muente-Kunigami, and C. Z.-W. Qiang. The global opportunity in IT-based services: Assessing and enhancing country competitiveness. Report, The World Bank, 2010. 70
1324. C. Sun, V. Le, Q. Zhang, and Z. Su. Toward understanding compiler bugs in GCC and LLVM. In *Proceedings of the 2016 International Symposium on Software Testing and Analysis*, ISSTA’16, pages 294–305, July 2016. 135
1325. S. Sun. What we are paying for: A quality adjusted price index for laptop microprocessors. Senior thesis, Wellesley College, Apr. 2014. 57, 61
1326. Sun Microsystems, Inc. Java code conventions. Technical report, Sun Microsystems, Inc, Sept. 1997. 171, 314
1327. T. Sunada, A. Monden, and K. Matsumoto. On estimating source lines of code from a binary program. In *Joint Conference of International Workshop on Software Measurement and International Conference on Software Process and Product Measurement*, IWSM/Mensura 2011, pages 3–6, Nov. 2011. 151
1328. K. Suzuki and S. Swanson. A survey of trends in non-volatile memory technologies: 2000-2014. In *IEEE International Memory Workshop*, IMW, pages 1–4, May 2015. 340

1329. T. N. Suzuki, D. Wheatcroft, and M. Griesser. Experimental evidence for compositional syntax in bird calls. *Nature Communications*, 7(10986), Mar. 2016. 16
1330. M. Swan and B. Smith. *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, second edition, 2001. 161
1331. G. M. Swift and S. M. Guertin. In-flight observations of multiple-bit upset in DRAMs. *IEEE Transactions on Nuclear Science*, 47(6):2386–2391, Dec. 2000. 142
1332. R. A. Syed, B. Robinson, and L. Williams. Does hardware configuration and processor load impact software fault observability? In *Third International Conference on Software Testing, Verification and Validation*, ICST, pages 285–294, Apr. 2010. 234, 235
1333. I. H. Tabernero. *A statistical examination of the properties and evolution of libre software*. PhD thesis, Universidad Rey Juan Carlos, Oct. 2008. 151, 239, 292, 293
1334. N. Taerat, N. Nakshinehaboon, C. Chandler, J. Elliot, C. B. Leangsuksun, G. Ostrouchov, S. L. Scott, and C. Englemann. Blue Gene/L log analysis and time to interrupt estimation. In *International Conference on Availability, Reliability and Security*, ARES'09, pages 173–180, Oct. 2009. 371
1335. P. P. Tallon, R. J. Kauffman, H. C. Lucas, A. B. Whinston, and K. Zhu. Using real options analysis for evaluating uncertain investments in information technology: Insights from the ICIS 2001 debate. *Communications of the Association for Information Systems*, 9:136–167, Sept. 2002. 101
1336. T. Tamai and Y. Torimitsu. Software lifetime and its evolution process over generations. In *Proceedings of 1992 Conference on Software Maintenance*, pages 63–69, Nov. 1992. 85
1337. P.-N. Tan and V. K. J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313, June 2004. 335
1338. E. Tang, E. Barr, X. Li, and Z. Su. Perturbing numerical calculations for statistical analysis of floating-point program (in)stability. In *Proceedings of the 19th international symposium on Software testing and analysis*, ISSTA'10, pages 131–142, July 2010. 125
1339. V. Tarasov, A. Mudrankit, W. Buik, P. Shilane, G. Kuennen, and E. Zadok. Generating realistic datasets for deduplication analysis. In *Proceedings of the 2012 USENIX Annual Technical Conference*, ATC'12, June 2012. 212, 321
1340. Q. C. Taylor. Analysis and characterization of author contribution patterns in open source software development. Thesis (m.s.), Brigham Young University, Apr. 2012. 57, 165
1341. M. Tedre. Computing as a science: A survey of competing viewpoints. *Minds & Machines*, 21(3):361–387, Aug. 2011. 5
1342. J. Teixeira, G. Robles, and J. M. González-Barahona. Lessons learned from applying social network analysis on an industrial free/libre/open source software ecosystem. *Journal of Internet Services and Applications*, 6(1):1–27, 2015. 121
1343. M. Templ, B. Meindl, and A. Kowarik. Introduction to statistical disclosure control (SDC). Technical report, International Household Survey Network, Oct. 2015. 364
1344. K. Tentori, D. Osherson, L. Hasher, and C. May. Wisdom and ageing: Irrational preferences in college students but not older adults. *Cognition*, 81(3):B87–B99, 2001. 44
1345. P. E. Tetlock. Accountability: The neglected social context of judgment and choice. *Research in Organizational Behavior*, 7:297–332, 1985. 47
1346. P. E. Tetlock. An alternative metaphor in the study of judgment and choice: People as politicians. *Theory and Psychology*, 1(4):451–475, 1991. 47
1347. P. E. Tetlock, O. V. Kristel, S. B. Elson, M. C. Green, and J. S. Lerner. The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, 78:853–870, 2000. 45
1348. Tezzaron Semiconductor. Soft errors in electronic memory. Technical Report 1.1, Tezzaron Semiconductor, Naperville, IL, Jan. 2004. 142
1349. T. A. Thayer, M. Lipow, and E. C. Nelson. *Software Reliability*. North-Holland Publishing Company, 1978. 5, 126
1350. E. Thereska, B. Doebel, A. X. Zheng, and P. Nobel. Practical performance models for complex, popular applications. In *Performance Evaluation Review*, SIGMETRICS'10, pages 1–12, June 2010. 185, 186, 346
1351. D. R. Thomas. *Security metrics for computer systems*. PhD thesis, Cambridge Computer Laboratory, University of Cambridge, Sept. 2015. 125, 129
1352. M. Thomas and V. Morwitz. Penny wise and pound foolish: The left-digit effect in price cognition. *Journal of Consumer Research*, 32(1):54–64, June 2005. 62
1353. M. Thomas, D. H. Simon, and V. Kadiyali. Do consumers perceive precise prices to be lower than round prices? Evidence from laboratory and market data. Research Paper Series #09-07, Johnson School, Cornell University, Sept. 2007. 62
1354. P. Thompson. How much did the Liberty shipbuilders forget? *Management Science*, 53(6):908–918, June 2007. 76
1355. S. Thummalapenta, L. Cerulo, L. Aversano, and M. D. Penta. An empirical study on the maintenance of source code clones. *Empirical Software Engineering*, 15(1):1–34, Feb. 2010. 6, 95
1356. M. A. Tinker. The relative legibility of the letters, the digits, and of certain mathematical signs. *Journal of Generative Psychology*, 1:472–494, 1928. 162
1357. J. T. Townsend. Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 9(1A):40–50, 1971. 42
1358. T. S. Traaen. The Brooks Act: An 8-bit act in a 64-bit world? An investigation of the Brooks Act and its implications to the department of defense information technology acquisition process. Executive Research Project S18, The Industrial College of the Armed Forces, National Defense University, Washington, D.C., May 1995. 76
1359. Transport, Department for. The accidents sub-objective. Transport Analysis Guidance Unit 3.4.1, Department for Transport, United Kingdom, Apr. 2011. 127, 128
1360. A penny saved: Psychological pricing. website, Oct. 2013. <http://blog.gumroad.com/post/64417917582/a-penny-saved-psychological-pricing>. 62
1361. A. Treisman and J. Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285–310, 1985. 33
1362. L. M. Trick and Z. W. Pylyshyn. What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, 19(2):331–351, 1993. 39, 40
1363. J. E. Triplett. Performance measures for computers. In *Deconstructing the Computer*, pages 99–139, Feb. 2003. 1
1364. K. S. Trivedi. *Probability & Statistics with Reliability, Queuing and Computer Science Applications*. John Wiley & Sons, Inc, second edition, 2002. 213
1365. C.-C. Tsai, B. Jain, N. A. Abdul, and D. E. Porter. A study of modern Linux API usage and compatibility: What to support when you're supporting. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys'16, page 16, Apr. 2016. 92
1366. N. P. Tschacher. Typosquatting in programming language package managers. Thesis (b.sc.), Department of Informatics, University of Hamburg, Mar. 2016. 137
1367. TSMC. TSMC historical operating data. [http://www.tsmc.com/english/investorRelations/historical\\_information.htm](http://www.tsmc.com/english/investorRelations/historical_information.htm), May 2017. 79
1368. T. S. Tullis and J. N. Stetson. A comparison of questionnaires for assessing website usability. In *Proceedings of Usability Professionals Association*, pages 1–12, June 2004. 345
1369. J. Turley. Embedded processors. <http://www.extremetech.com>, Jan. 2002. 78, 256
1370. H. Turner and D. Firth. Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, 48(9):1–21, 2012. 314
1371. H. Turner and D. Firth. *Generalized nonlinear models in R: An overview of the gnm package*. University of Warwick, UK, 1.0-8 edition, Apr. 2015. 276
1372. M. L. Turner and R. W. Engle. Is working memory capacity task dependent? *Journal of Memory and Language*, 28(2):127–154, Apr. 1989. 158
1373. R. Turner. *Weathering Heights: The Emergence of Aeronautical Meteorology as an Infrastructural Science*. PhD thesis, University of Pennsylvania, May 2010. 2, 67

1374. J. Tzelgov, V. Yehene, L. Kotler, and A. Alon. Automatic comparisons of artificial digits never compared: Learning linear ordering relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1):103–120, 2000. [41](#)
1375. A. Čaušević, R. Shukla, S. Punnekkat, and D. Sundmark. Effects of negative testing on TDD: An industrial experiment. In H. Baumeister and B. Weber, editors, *Agile Processes in Software Engineering and Extreme Programming*, volume 149 of *Lecture Notes in Business Information Processing*, pages 91–105. Springer Berlin Heidelberg, 2013. [147](#)
1376. The ultimate Debian database. website, 2014. <http://wiki.debian.org/UltimateDebianDatabase/>. [126](#), [182](#), [253](#)
1377. G. Ülkümen, M. Thomas, and V. G. Morwitz. Will i spend more in 12 months or a year? The effect of ease of estimation and confidence on budget estimates. *Journal of Consumer Research*, 35(2):245–256, Aug. 2008. [110](#)
1378. Unicode Consortium, The. *The Unicode Standard Version 11.0 – Core Specification*. The Unicode Consortium, June 2018. [72](#)
1379. Defence technical information center. Search page for DTIC reports, July 2016. <http://dsearch.dtic.mil>. [5](#)
1380. D. Šmite, R. Britto, and R. van Solingen. Calculating the extra costs and the bottom-line hourly cost of offshoring. In *2017 IEEE 12th International Conference on Global Software Engineering*, ICGSE, page ???, May 2017. [108](#)
1381. I. Utting, D. Bouvier, M. Caspersen, A. E. Tew, R. Frye, Y. B.-D. Kolikant, M. McCracken, J. Paterson, J. Sorva, L. Thomas, and T. Wilusz. A fresh look at novice programmers’ performance and their teachers’ expectations. In *ITiCSE-WGR’13*, pages 15–32, June 2013. [320](#)
1382. Ž Antolić. Fault slip through measurement process implementation in CPP software verification. In *International Conference on Business Intelligence Systems*, miproBIS 2007, page ???, May 2007. [144](#)
1383. J. v. Kistowski, H. Block, J. Beckett, K.-D. Lange, J. A. Arnold, and S. Kounev. Analysis of the influences on server power consumption and energy efficiency for CPU-intensive workloads. In *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, ICPE’15, pages 223–234, Jan. 2015. [221](#)
1384. A. Vahabzadeh, A. M. Fard, and A. Mesbah. An empirical study of bugs in test code. In *International Conference on Software Maintenance and Evolution*, ICSME 2015, pages 101–110, Oct. 2015. [146](#)
1385. O. Van den Bergh, S. Vrana, and P. Eelen. Letters from the heart: Affective categorization of letter combinations in typists and nontypists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6):1153–1161, 1990. [73](#)
1386. K. G. van den Boogaart and R. Tolosana-Delgado. *Analyzing Compositional Data with R*. Springer, 2013. [307](#), [308](#)
1387. E. van der Kouwe, D. Andriesse, H. Bos, C. Giuffrida, and G. Heiser. Benchmarking crimes: An emerging threat in systems security. In *eprint arXiv:cs.CR/1801.02381*, Jan. 2018. [336](#)
1388. C. van der Merwe. An engineering approach to an integrated value proposition design framework. Thesis (m.s.), Faculty of Industrial Engineering at Stellenbosch University, Mar. 2015. [63](#)
1389. M. J. P. van der Meulen. *The Effectiveness of Software Diversity*. PhD thesis, Centre for Software Reliability, City University, Nov. 2007. [109](#), [150](#), [153](#), [207](#)
1390. M. J. P. van der Meulen, P. G. Bishop, and M. Revilla. An exploration of software faults and failure behaviour in a large population of programs. In *15th International Symposium on Software Reliability Engineering*, ISSRE 2004, pages 101–120, Nov. 2004. [138](#)
1391. M. P. van Oeffelen and P. G. Vos. A probabilistic model for the discrimination of visual number. *Perception & Psychophysics*, 32(2):163–170, 1982. [40](#)
1392. K. E. van Oorschot, J. W. M. Bertrand, and C. G. Rutte. Field studies into the dynamics of product development tasks. *International Journal of Operations & Production Management*, 25(8):720–739, 2005. [116](#)
1393. H. VanLehn. *Mind Bugs: The Origins of Procedural Misconceptions*. The MIT Press, 1990. [40](#), [41](#)
1394. R. Vasa. *Growth and Change Dynamics in Open Source Software Systems*. PhD thesis, Faculty of Information and Communication Technology, Swinburne University of Technology, Melbourne, Oct. 2010. [83](#), [233](#), [247](#)
1395. B. Vasilescu, A. Serebrenik, M. Goeminne, and T. Mens. On the variation and specialisation of workload-A case study of the Gnome ecosystem community. *Empirical Software Engineering*, 19(4):955–1008, Aug. 2012. [254](#)
1396. P. Vassiliadis, M.-R. Kolozoff, M. Zerva, and A. V. Zarras. Schema evolution and foreign keys: Birth, eviction, change and absence. In *International Conference on Conceptual Modeling*, ER 2017, pages 106–119, Nov. 2017. [96](#)
1397. VCDB. VERIS community database. <https://github.com/vz-risk/VCDB>, Mar. 2018. [126](#)
1398. W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, fourth edition, 2002. [292](#)
1399. P. Verghese and D. G. Pelli. The information capacity of visual attention. *Vision Research*, 32(5):983–995, 1992. [50](#)
1400. C. Verhoef. Quantitative IT portfolio management. *Science of Computer Programming*, 45(1):1–96, Oct. 2002. [107](#)
1401. I. Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2):219–240, Mar. 1991. [188](#)
1402. A. Vetro, N. Zazwarka, C. Seaman, and F. Shull. Using the ISO/IEC 9126 product quality model to classify defects: a controlled experiment. In *16th International Conference on Evaluation & Assessment in Software Engineering*, EASE 2012, pages 187–196, May 2012. [127](#), [282](#)
1403. B. Veysman and L. Akhmadeeva. Towards evidence-based typography: First results. *TUGboat*, 33(2):156–156, Apr. 2012. [204](#)
1404. Vgchartz global yearly chart: 2005-2016. website, Feb. 2017. <http://www.vgchartz.com/yearly/2016/Global/>. [63](#)
1405. V. B. Viard. Information goods upgrades: Theory and evidence. *The B.E. Journal of Theoretical Economics*, 7(1):1–34, 2007. [64](#)
1406. N. M. Victor and J. H. Ausubel. DRAMs as model organisms for study of technological evolution. *Technological Forecasting and Social Change*, 69(3):243–262, Apr. 2002. [67](#)
1407. S. Vidal, A. Bergel, J. A. Díaz-Pace, and C. Marcosa. Over-exposed classes in Java: An empirical study. *Computer Languages, Systems & Structures*, 46:1–19, Nov. 2016. [171](#)
1408. List of most expensive video games to develop. website, 2018. [https://en.wikipedia.org/List\\_of\\_most\\_expensive\\_video\\_games\\_to\\_develop](https://en.wikipedia.org>List_of_most_expensive_video_games_to_develop). [54](#)
1409. F. Viénot, H. Brettel, and J. D. Mollon. Digital video colourmaps for checking the legibility of displays by dichromats. *COLOR research and application*, 24(4):243–252, Aug. 1999. [193](#)
1410. V. Villard. Android version distribution history. <http://www.bidouille.org/misc/androidcharts>, 2015. [87](#), [88](#), [194](#)
1411. T. H. Vines, A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Renison. The availability of research data declines rapidly with article age. In *eprint arXiv:abs/1312.5670*, Dec. 2013. [7](#)
1412. W. K. Viscusi and J. E. Aldy. The value of a statistical life: A critical review of market estimates throughout the world. Working Paper No. 9487, National Bureau of Economic Research, USA, Feb. 2003. [127](#)
1413. K. D. Vogeeler, G. Memmi, and P. Jouvelot. Parameter sensitivity analysis of the energy/frequency convexity rule for nanometer-scale application processors. In *eprint arXiv:cs.DS/1508.07740*, Aug. 2015. [338](#)
1414. K. von Fintel and L. Matthewson. Universals in semantics. *The Linguistic Review*, 25(1-2):139–201, 2008. [30](#)
1415. S. L. R. Vrhovec, T. Hovelja, D. Vavpotič, and M. Krisper. Diagnosing organizational risks in software projects: Stakeholder resistance. *International Journal of Project Management*, 33(6):1262–1273, Aug. 2015. [117](#)
1416. M. Wachs. When planners lie with numbers. *Journal of the American Planning Association*, 55(4):476–479, Apr. 1989. [105](#)
1417. J. Wagemans, J. H. Elder, M. Kubovy, M. A. Peterson, S. E. Palmer, M. Singh, and R. von der Heydt. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin*, 138(6):1172–1217, 2012. [32](#)
1418. J. Wainer, C. G. N. Barsottini, D. Lacerda, and L. R. M. de Marco. Empirical evaluation in computer science research published by ACM. *Information and Software Technology*, 51(6):1081–1085, June 2009. [5](#)

1419. L. Wakeham. Government policy on the management of risk, volume I: Report. HL Paper 183-I, Select Committee on Economic Affairs, UK House of Lords, June 2006. 128
1420. S. Waligora, J. Bailey, and M. Stark. Impact of Ada and object-oriented design in the flight dynamics division at Goddard space flight center. Technical Report SEL-95-001, Goddard Space Flight Center, Mar. 1995. 153, 177
1421. D. R. Wallace and D. R. Kuhn. Failure modes in medical device software: An analysis of 15 years of recall data. *International Journal of Reliability, Quality and Safety Engineering*, 8(4):351–372, Dec. 2001. 126
1422. P. Wang. Chasing the hottest IT: Effects of information technology fashion on organizations. *MIS Quarterly*, 34(1):63–85, Mar. 2010. 4, 7
1423. W. Wang. Toward improved understanding and management of software clones. Thesis (m.s.), University of Waterloo, Ontario, Canada, May 2012. 94, 95
1424. Y. Wang. Language matters. In *Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM’15, page ???, Oct. 2015. 73
1425. Y. Wang and J. Zhang. The effort distribution of software development phases. *Computer Science and Application*, 7(5):428–437, May 2017. 106
1426. L. Wanner, C. Apte, R. Balani, P. Gupta, and M. Srivastava. A case for opportunistic embedded sensing in presence of hardware power variability. In *Proceedings of the 2010 international conference on Power aware computing and systems*, HotPower’10, pages 1–8, Oct. 2010. 339
1427. G. Ward, L. Tan, and R. Grenfell-Essam. Examining the relationship between free recall and immediate serial recall: the effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 36(5):1207–1241, Sept. 2010. 24
1428. C. Ware. *Information Visualization Perception for Design*. Morgan Kaufmann Publishers, 2000. 32
1429. W. H. Ware, S. N. Alexander, P. Armer, M. M. Astrahan, L. Bers, H. H. Goode, H. D. Huskey, and M. Rubinoff. Soviet computer technology—1959. Research Memorandum RM-2541, The RAND Corporation, Mar. 1960. 4
1430. P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, XII(3):129–140, 1960. 52
1431. P. C. Wason. Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20(3):273–281, 1968. 35
1432. J. Waters. Variable marginal propensities to pirate and the diffusion of computer software. MPRA Paper No. 46036, Nottingham University Business School, Apr. 2013. 63
1433. V. M. Weaver and J. Dongarra. Can hardware performance counters produce expected, deterministic results? In *3rd Workshop on Functionality of Hardware Performance Monitoring*, pages 1–11, Dec. 2010. 340
1434. V. M. Weaver and S. A. McKee. Can hardware performance counters be trusted? In *IEEE International Symposium on Workload Characterization*, IISWC’08, pages 141–150, Sept. 2008. 340
1435. E. U. Weber, A.-R. Blais, and N. E. Betz. A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavior and Decision Making*, 15(4):263–290, Apr. 2002. 44
1436. B. F. Webster. Patterns in IT litigation: Systems failure (1976-2000). A study, PriceaterhouseCoopers LLP, 2000. 111
1437. B. S. Weekes. Differential effects of number of letters on word and nonword naming latency. *The Quarterly Journal of Experimental Psychology*, 50A(2):439–456, 1997. 22
1438. D. M. Wegner. *The Illusion of Conscious Will*. MIT Press, 2002. 16
1439. M. H. Weik. A survey of domestic electronic digital computing systems. Technical Report 971, Ballistic Research Laboratories, Maryland, Dec. 1955. 336
1440. M. H. Weik. A third survey of domestic electronic digital computing systems. Technical Report 1115, Ballistic Research Laboratories, Maryland, Mar. 1961. 336
1441. G. F. Weinwurm and H. J. Zagorski. Research into the management of computer programming: A transitional analysis of cost estimation techniques. Technical Documentary Report ESD-TR-65-575, United States Air Force, L. G. Hanscom Field, Bedford, Massachusetts, Nov. 1965. 106
1442. J. A. White. Grapher pics. <http://www.talljerome.com/mathnerd.html>, Oct. 2012. 196
1443. M. White. Scaled CMOS technology reliability users guide. JPL Publication 09-33 01/10, Jet Propulsion Laboratory, California Institute of Technology, 2010. 141
1444. White House, The. Guidelines and discount rates for benefit-cost analysis of federal programs. OMB Circular A-94, US Government, 1992. 56
1445. D. Whitfield. Cost overruns, delays and terminations: 105 outsourced public sector ICT projects. ESSU Research Report 3, European Services Strategy Unit, Dec. 2007. 104, 105
1446. R. M. Whyte. Order Re Sun’s Motions for Preliminary Injunction Against Microsoft. Re: Sun Microsystems v. Microsoft, Case No. 97-20884 RMW(PVT). Opinion, UNITED STATES DISTRICT COURT FOR THE NORTHERN DISTRICT OF CALIFORNIA, 1998. 70
1447. J. M. Wicherts, M. Bakker, and D. Molenaar. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE*, 6(11):e26828, Nov. 2011. 7
1448. W. A. Wickelgren. Size of rehearsal group and short-term memory. *Journal of Experimental Psychology*, 68(4):413–419, 1964. 24
1449. G. Wiederhold. What is your software worth? Technical Report ???, Stanford University, Apr. 2007. 60
1450. A. Wierzbicka. *Semantics: Primes and Universals*. Oxford University Press, 1996. 30
1451. R. Wilcox. *Introduction to Robust Estimation & Hypothesis Testing*. Elsevier, 3rd edition, 2012. 227
1452. J. Wiley. Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & Cognition*, 26(4):716–730, 1998. 31, 75
1453. M. V. Wilkes. *Memoirs of a Computer Pioneer*. MIT Press, 1984. 123
1454. M. V. Wilkes, D. J. Wheeler, and S. Gill. *The Preparation of Programs for an Electronic Digital Computer*. Addison-Wesley Publishing Company, Inc., second edition, 1957. 91
1455. L. Wilkinson. *The Grammar of Graphics*. Springer, second edition, 2005. 189
1456. P. Williams and B. Curtis. A matched project evaluation of modern programming practices: Scientific report on the ASTROS plan. Technical Report RADC-TR-80-6, Vol II, General Electric Company, Feb. 1980. 121
1457. R. R. Willis, R. M. Rova, M. D. Scott, M. I. Johnson, J. F. Ryskowski, J. A. Moon, K. C. Shumate, and T. O. Winfield. Hughes Aircraft’s widespread deployment of a continuously improving software process. Technical Report CMU/SEI-98-TR-006, Raytheon Systems Company, May 1998. 60
1458. H. E. Willman, Jr., T. A. James, A. A. Beauregard, and P. Hilcoff. Software systems reliability: A Raytheon project history. Final Technical Report RADC-TR-77-188, Rome Air Development Center, June 1977. 126
1459. L. M. Wills. Automated program recognition by graph parsing. A.I. Technical Report No. 1358, MIT Artificial Intelligence Laboratory, July 1992. 168
1460. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, second edition, 2005. 310
1461. R. W. Wolverton. The cost of developing large-scale software. *IEEE Transactions on Computers*, c-23(6):615–636, June 1974. 113
1462. W. E. Wong, S. S. Gokhale, and J. R. Horgan. Quantifying the closeness between program components and features. *The Journal of Systems and Software*, 54(2):87–98, Oct. 2000. 138, 168
1463. A. Wood. Software reliability growth models. Technical Report 96.1, Tandem Computer, Sept. 1996. 132, 133
1464. Semiconductor monthly sales volume: 1975-2016. website, Mar. 2016. <https://www.wsts.org>. 4
1465. D. Wren. Passmark website??? <http://www.passmark.com>, July 2014. 344, 345
1466. R. Wright. *The Evolution of GOD*. Little, Brown Book Group, 2009. 68

1467. Microsoft server protocol documentation. website, 2015. <http://www.microsoft.com>. 80, 88, 184, 185
1468. S. D. Wu, C. Rossin, K. G. Kempf, M. O. Atan, B. Aytac, S. A. Shirodkar, and A. Mishra. Extending Bass for improved new product forecasting. ???, ???(??):???, Apr. 2009. 63
1469. J. Yan and W. Zhang. Compiler-guided register reliability improvement against soft errors. In *Proceedings of the 5th ACM international conference on Embedded software*, EMSOFT'05, pages 203–209, Sept. 2005. 142
1470. M. Yang, G.-R. Uh, and D. B. Whalley. Efficient and effective branch reordering using profile data. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 24(6):667–697, Nov. 2002. 169
1471. M. C. K. Yang and A. Chao. Reliability-estimation & stopping-rules for software testing, based on repeated appearances of bugs. *IEEE Transactions on Reliability*, 44(2):315–321, June 1995. 148
1472. X. Yang, Z. Wang, J. Xue, and Y. Zhou. The reliability wall for exascale supercomputing. *IEEE Transactions on Computers*, 61(6):767–779, June 2011. 143
1473. M. J. Yap, S. J. R. Liow, S. B. Jalil, and S. S. B. Faizal. The Malay lexicon project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42(4):992–1003, Nov. 2010. 161
1474. Y. C. B. Yeh. Triple-triple redundant 777 primary flight computer. In *Proceedings Aerospace Applications Conference (vol 1)*, pages 293–307, Feb. 1996. 142
1475. J. R. Yost. *Making IT Work: A History of the Computer Services Industry*. The MIT Press, 2017. 77
1476. A. G. Yu. *Managing Application Software Suppliers in Information System Development Projects*. PhD thesis, Department of Management and Organisation, University of Stirling, Nov. 2003. 112
1477. L. Yu and S. Ramaswamy. A study of SourceForge users and user network. AAAI Technical Report FS-13-05, Association for the Advancement of Artificial Intelligence, Nov. 2013. 167
1478. D. Yuan, S. Park, and Y. Zhou. Characterizing logging practices in open-source software. In *Proceedings of the 34th International Conference on Software Engineering*, ICSE 2012, pages 102–112, June 2012. 143
1479. T. Yuki and S. Rajopadhye. Folklore confirmed: Compiling for speed = compiling for energy. Technical Report CS13-107, Computer Science Department, Colorado State University, Aug. 2013. 338
1480. A. Zaidman, B. V. Rompaey, A. van Deursen, and S. Demeyer. Studying the co-evolution of production and test code in open source and industrial developer test processes through repository mining. Technical Report TUD-SERG-2010-035, Software Engineering Research Group, Delft University of Technology, 2010. 146
1481. S. F. Zeigler. Comparing development costs of C and Ada. Technical report, Rational Software Corporation, Mar. 1995. 51
1482. A. Zeileis, K. Hornik, and P. Murrell. Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis*, 53(9):3259–3270, July 2009. 193
1483. M. V. Zelkowitz. The effectiveness of software prototyping: A case study. In *26th Annual Technical Symposium ???*, pages 7–15, June 1987. 116
1484. A. Zeller, T. Zimmermann, and C. Bird. Failure is a four-letter word—A parody in empirical research-. In *Proceedings of the 7th International Conference on Predictive Models in Software Engineering*, PROMISE'11, pages 5:1–5:7, Sept. 2011. 238
1485. J. Zhang and H. Wang. The effect of external representations on numeric tasks. *The Quarterly Journal of Experimental Psychology*, 58(5):817–838, Oct. 2005. 40
1486. J. Zhang, M. Zhu, D. Hao, and L. Zhang. An empirical study on the scalability of selective mutation testing. In *2014 IEEE 25th International Symposium on Software Reliability Engineering*, pages 277–287, Nov. 2014. 148
1487. X. Zhang. *An Analysis of the Effect of Environmental and Systems Complexity on Information Systems Failures*. PhD thesis, University of North Texas, Aug. 2001. 69
1488. Y. Zhang, Y. Jiang, C. Xu, X. Ma, and P. Yu. ABC: Accelerated building of C/C++ projects. In *Asia-Pacific Software Engineering Conference*, APSEC 2015, pages 182–189, Dec. 2015. 172
1489. Y. Zhang, J. W. Lee, N. P. Johnson, and D. I. August. DAFT: Decoupled acyclic fault tolerance. In *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*, PACT'10, pages 87–98, Sept. 2010. 142
1490. M. Zhao, J. Grossklags, and P. Liu. An empirical study of web vulnerability discovery ecosystems. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS'15, pages 1105–1117, Oct. 2015. 71, 126
1491. M. Zhao and P. Liu. Empirical analysis and modeling of black-box mutational fuzzing. In *International Symposium on Engineering Secure Software and Systems*, ESSoS 2016, pages 173–189, Apr. 2016. 133, 134
1492. Y. Zhao, A. Serebrenik, Y. Zhou, V. Filkov, and B. Vasilescu. The impact of continuous integration on other software development practices: A large-scale empirical study. In *Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering*, ASE 2017, pages 60–71, Oct.-Nov. 2017. 120
1493. J. Zheng, L. Williams, N. Nagappan, W. Snipes, J. P. Hudepohl, and M. A. Vouk. On the value of static analysis for fault detection in software. *IEEE Transactions on Software Engineering*, 32(4):240–253, Apr. 2006. 145
1494. H. Zhong and Z. Su. An empirical study on real bug fixes. In *Proceedings of the 37th International Conference on Software Engineering*, ICSE'15, pages 913–923, May 2015. 140
1495. H. Zhou and A. Fishbach. The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, 111(4):493–504, Oct. 2016. 321
1496. K. Zhou, P. Huang, C. Li, and H. Wang. An empirical study on the interplay between filesystems and SSD. In *7th International Conference on Networking, Architecture and Storage*, NAS, pages 124–133, June 2012. 343
1497. X. Zhu, E. J. Whitehead, Jr., C. Sadowski, and Q. Song. An analysis of programming language statement frequency in C, C++, and Java source code. *Software—Practice and Experience*, 15(11):1479–1495, Nov. 2015. 205, 206
1498. A. Ziegler, V. Rothberg, and D. Lohmann. Analyzing the impact of feature changes in Linux. In *Proceedings of the Tenth International Workshop on Variability Modelling of Software-intensive Systems*, VaMoS'16, pages 25–32, Jan. 2016. 168
1499. T. Zimmermann, N. Nagappan, H. Gall, E. Giger, and B. Murphy. Cross-project defect prediction. In *Proceedings of the 7th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering*, ESEC/FSE 2009, pages 91–100, Aug. 2009. 144
1500. T. Zimmermann, R. Premraj, and A. Zeller. Predicting defects for Eclipse. In *Proceedings of the Third International Workshop on Predictor Models in Software Engineering*, PROMISE'07, May 2007. 140
1501. P. M. Zislis. An experiment in algorithm implementation. Technical Report CSD-TR-96, Purdue University, Indiana, USA, June 1973. 25, 153, 164
1502. F. Zlotnick. *The POSIX.1 Standard: A Programmer's Guide*. The Benjamin/Cummings Publishing Company, 1991. 91
1503. K. Zuse. Über den allgemeinen Plankalkül als mittel zur formulierung schematisch-kombinatoriver aufgaben. *Archiv der Mathematik*, 1:441–449, 1949. 89
1504. O. Zwikel and S. Globerson. Benchmarking of project planning and success in selected industries. *Benchmarking: An International Journal*, 13(6):688–700, 2006. 102