

INTRODUCTION TO  
PROBABILITY THEORY  
AND STATISTICS  
FOR PSYCHOLOGY

AND

QUANTITATIVE METHODS FOR  
HUMAN SCIENCES

David Steinsaltz<sup>1</sup>  
University of Oxford  
(Lectures 1–8 based on earlier version by Jonathan Marchini)

Lectures 1–8: MT 2011  
Lectures 9–16: HT 2012

---

<sup>1</sup>University lecturer at the Department of Statistics, University of Oxford



# Contents

<b>1</b>	<b>Describing Data</b>	<b>1</b>
1.1	Example: Designing experiments . . . . .	1
1.2	Variables . . . . .	2
1.2.1	Types of variables . . . . .	2
1.2.2	Ambiguous data types . . . . .	3
1.3	Plotting Data . . . . .	5
1.3.1	Bar Charts . . . . .	6
1.3.2	Histograms . . . . .	6
1.3.3	Cumulative and Relative Cumulative Frequency Plots and Curves . . . . .	8
1.3.4	Dot plot . . . . .	11
1.3.5	Scatter Plots . . . . .	12
1.3.6	Box Plots . . . . .	12
1.4	Summary Measures . . . . .	12
1.4.1	Measures of location (Measuring the center point) . .	13
1.4.2	Measures of dispersion (Measuring the spread) . . .	17
1.5	Box Plots . . . . .	21
1.6	Appendix . . . . .	22
1.6.1	Mathematical notation for variables and samples . .	22
1.6.2	Summation notation . . . . .	23
<b>2</b>	<b>Probability I</b>	<b>25</b>
2.1	Why do we need to learn about probability? . . . . .	25
2.2	What is probability? . . . . .	27
2.2.1	Definitions . . . . .	28
2.2.2	Calculating simple probabilities . . . . .	28
2.2.3	Example 2.3 continued . . . . .	28
2.2.4	Intersection . . . . .	29
2.2.5	Union . . . . .	29

2.2.6	Complement . . . . .	30
2.3	Probability in more general settings . . . . .	30
2.3.1	Probability Axioms (Building Blocks) . . . . .	31
2.3.2	Complement Law . . . . .	31
2.3.3	Addition Law (Union) . . . . .	31
<b>3</b>	<b>Probability II</b>	<b>33</b>
3.1	Independence and the Multiplication Law . . . . .	33
3.2	Conditional Probability Laws . . . . .	36
3.2.1	Independence of Events . . . . .	37
3.2.2	The Partition law . . . . .	38
3.3	Bayes' Rule . . . . .	39
3.4	Probability Laws . . . . .	41
3.5	Permutations and Combinations (Probabilities of patterns) .	41
3.5.1	Permutations of $n$ objects . . . . .	41
3.5.2	Permutations of $r$ objects from $n$ . . . . .	42
3.5.3	Combinations of $r$ objects from $n$ . . . . .	44
3.6	Worked Examples . . . . .	45
<b>4</b>	<b>The Binomial Distribution</b>	<b>49</b>
4.1	Introduction . . . . .	49
4.2	An example of the Binomial distribution . . . . .	49
4.3	The Binomial distribution . . . . .	51
4.4	The mean and variance of the Binomial distribution . . . . .	52
4.5	Testing a hypothesis using the Binomial distribution . . . . .	53
<b>5</b>	<b>The Poisson Distribution</b>	<b>57</b>
5.1	Introduction . . . . .	57
5.2	The Poisson Distribution . . . . .	59
5.3	The shape of the Poisson distribution . . . . .	61
5.4	Mean and Variance of the Poisson distribution . . . . .	61
5.5	Changing the size of the interval . . . . .	62
5.6	Sum of two Poisson variables . . . . .	62
5.7	Fitting a Poisson distribution . . . . .	63
5.8	Using the Poisson to approximate the Binomial . . . . .	64
5.9	Derivation of the Poisson distribution (non-examinable) . . .	67
5.9.1	Error bounds (very mathematical) . . . . .	67

<b>6 The Normal Distribution</b>	<b>69</b>
6.1 Introduction . . . . .	69
6.2 Continuous probability distributions . . . . .	71
6.3 What is the Normal Distribution? . . . . .	71
6.4 Using the Normal table . . . . .	72
6.5 Standardisation . . . . .	76
6.6 Linear combinations of Normal random variables . . . . .	78
6.7 Using the Normal tables backwards . . . . .	81
6.8 The Normal approximation to the Binomial . . . . .	82
6.8.1 Continuity correction . . . . .	82
6.9 The Normal approximation to the Poisson . . . . .	83
<b>7 Confidence intervals and Normal Approximation</b>	<b>87</b>
7.1 Confidence interval for sampling from a normally distributed population . . . . .	87
7.2 Interpreting the confidence interval . . . . .	88
7.3 Confidence intervals for probability of success . . . . .	91
7.4 The Normal Approximation . . . . .	91
7.4.1 Normal distribution . . . . .	92
7.4.2 Poisson distribution . . . . .	92
7.4.3 Bernoulli variables . . . . .	92
7.5 CLT for real data . . . . .	94
7.5.1 Quebec births . . . . .	94
7.5.2 California incomes . . . . .	97
7.6 Using the Normal approximation for statistical inference . . . . .	97
7.6.1 An example: Average incomes . . . . .	97
<b>8 The Z Test</b>	<b>101</b>
8.1 Introduction . . . . .	101
8.2 The logic of significance tests . . . . .	101
8.2.1 Outline of significance tests . . . . .	103
8.2.2 Significance tests or hypothesis tests? Breaking the .05 barrier . . . . .	103
8.2.3 Overview of Hypothesis Testing . . . . .	104
8.3 The one-sample Z test . . . . .	105
8.3.1 Test for a population mean $\mu$ . . . . .	106
8.3.2 Test for a sum . . . . .	106
8.3.3 Test for a total number of successes . . . . .	106
8.3.4 Test for a proportion . . . . .	107
8.3.5 General principles: The square-root law . . . . .	109

8.4 One and two-tailed tests . . . . .	109
8.5 Hypothesis tests and confidence intervals . . . . .	109
<b>9 The <math>\chi^2</math> Test</b>	<b>111</b>
9.1 Introduction — Test statistics that aren't Z . . . . .	111
9.2 Goodness-of-Fit Tests . . . . .	112
9.2.1 The $\chi^2$ distribution . . . . .	113
9.2.2 Large d.f. . . . .	115
9.3 Fixed distributions . . . . .	116
9.4 Families of distributions . . . . .	119
9.4.1 The Poisson Distribution . . . . .	119
9.4.2 The Binomial Distribution . . . . .	120
9.5 Chi-squared Tests of Association . . . . .	122
<b>10 The T distribution and Introduction to Sampling</b>	<b>125</b>
10.1 Using the T distribution . . . . .	125
10.1.1 Using t for confidence intervals: Single sample . . . . .	126
10.1.2 Using the T table . . . . .	127
10.1.3 Using t for Hypothesis tests . . . . .	128
10.1.4 When do you use the Z or the T statistics? . . . . .	129
10.1.5 Why do we divide by $n - 1$ in computing the sample SD? . . . . .	129
10.2 Paired-sample t test . . . . .	130
10.3 Introduction to sampling . . . . .	130
10.3.1 Sampling with and without replacement . . . . .	130
10.3.2 Measurement bias . . . . .	131
10.3.3 Bias in surveys . . . . .	131
10.3.4 Measurement error . . . . .	133
<b>11 Comparing Distributions</b>	<b>135</b>
11.1 Normal confidence interval for difference between two population means . . . . .	135
11.2 Z test for the difference between population means . . . . .	136
11.3 Z test for the difference between proportions . . . . .	136
11.4 t confidence interval for the difference between population means . . . . .	137
11.5 Two-sample test and paired-sample test. . . . .	138
11.5.1 Schizophrenia study: Two-sample t test . . . . .	138
11.5.2 The paired-sample test . . . . .	138
11.5.3 Is the CLT justified? . . . . .	139

11.6 Hypothesis tests for experiments . . . . .	140
11.6.1 Quantitative experiments . . . . .	140
11.6.2 Qualitative experiments . . . . .	141
<b>12 Non-Parametric Tests, Part I</b>	<b>143</b>
12.1 Introduction: Why do we need distribution-free tests? . . . . .	143
12.2 First example: Learning to Walk . . . . .	143
12.2.1 A first attempt . . . . .	143
12.2.2 What could go wrong with the T test? . . . . .	144
12.2.3 How much does the non-normality matter? . . . . .	144
12.3 Tests for independent samples . . . . .	147
12.3.1 Median test . . . . .	147
12.3.2 Rank-Sum test . . . . .	149
12.4 Tests for paired data . . . . .	150
12.4.1 Sign test . . . . .	150
12.4.2 Breastfeeding study . . . . .	151
12.4.3 Wilcoxon signed-rank test . . . . .	152
12.4.4 The logic of non-parametric tests . . . . .	153
<b>13 Non-Parametric Tests Part II, Power of Tests</b>	<b>155</b>
13.1 Kolmogorov-Smirnov Test . . . . .	155
13.1.1 Comparing a single sample to a distribution . . . . .	155
13.1.2 Comparing two samples: Continuous distributions . .	160
13.1.3 Comparing two samples: Discrete samples . . . . .	161
13.1.4 Comparing tests to compare distributions . . . . .	162
13.2 Power of a test . . . . .	162
13.2.1 Computing power . . . . .	162
13.2.2 Computing trial sizes . . . . .	163
13.2.3 Power and non-parametric tests . . . . .	164
<b>14 ANOVA and the F test</b>	<b>167</b>
14.1 Example: Breastfeeding and intelligence . . . . .	167
14.2 Digression: Confounding and the “adjusted means” . . .	167
14.3 Multiple comparisons . . . . .	168
14.3.1 Discretisation and the $\chi^2$ test . . . . .	168
14.3.2 Multiple t tests . . . . .	168
14.4 The F test . . . . .	169
14.4.1 General approach . . . . .	169
14.4.2 The breastfeeding study: ANOVA analysis . . . . .	172
14.4.3 Another Example: Exercising rats . . . . .	172

14.5 Multifactor ANOVA . . . . .	173
14.6 Kruskal-Wallis Test . . . . .	174
<b>15 Regression and correlation: Detecting trends</b>	<b>177</b>
15.1 Introduction: Linear relationships between variables . . . . .	177
15.2 Scatterplots . . . . .	178
15.3 Correlation: Definition and interpretation . . . . .	179
15.4 Computing correlation . . . . .	180
15.4.1 Brain measurements and IQ . . . . .	180
15.4.2 Galton parent-child data . . . . .	182
15.4.3 Breastfeeding example . . . . .	182
15.5 Testing correlation . . . . .	184
15.6 The regression line . . . . .	184
15.6.1 The SD line . . . . .	184
15.6.2 The regression line . . . . .	185
15.6.3 Confidence interval for the slope . . . . .	187
15.6.4 Example: Brain measurements . . . . .	189
<b>16 Regression, Continued</b>	<b>193</b>
16.1 $R^2$ . . . . .	193
16.1.1 Example: Parent-Child heights . . . . .	193
16.1.2 Example: Breastfeeding and IQ . . . . .	193
16.2 Regression to the mean and the regression fallacy . . . . .	195
16.3 When the data don't fit the model . . . . .	196
16.3.1 Transforming the data . . . . .	196
16.3.2 Spearman's Rank Correlation Coefficient . . . . .	196
16.3.3 Computing Spearman's rank correlation coefficient .	197

# Lecture 1

## Describing Data

*Uncertain knowledge  
+ knowledge about the extent of uncertainty in it  
= Useable knowledge*

C. R. Rao, statistician

*As we know, there are known knowns. There are things we know we know. We also know there are known unknowns. That is to say we know there are some things we do not know. But there are also unknown unknowns, The ones we don't know we don't know.*

Donald Rumsfeld, US Secretary of defense

Observations and measurements are at the centre of modern science. We put our ideas to the test by comparing them to what we find out in the world. Easier said than done, because all observation and measurement is uncertain. Some of the reasons are:

**Sampling** Our observations are only a small sample of the range of possible observations — the *population*.

**Errors** Every measurement suffers from errors.

**Complexity** The more observations we have, the more difficult it becomes to make them tell a coherent story.

We can never observe everything, nor can we make measures without error. But, as the quotes above suggest, uncertainty is not such a problem if it can be constrained — that is, if we know the limits of our uncertainty.

## 1.1 Example: Designing experiments

“If a newborn infant is held under his arms and his bare feet are permitted to touch a flat surface, he will perform well-coordinated walking movements similar to those of an adult[...] Normally, the walking and placing reflexes disappear by about 8 weeks.” [ZZK72] The question is raised, whether exercising this reflex would enable children to more quickly acquire the ability to walk independently. How would we resolve this question?

Of course, we could perform an experiment. We could do these exercises with an infant, starting from when he or she was a newborn, and follow up every week for about a year, to find out when this baby starts walking. Suppose it is 10 months. Have we answered the question then?

The obvious problem, then, is that we don’t know what age this baby would have started walking at without exercise. One solution would be to take another infant, observe this one at the same weekly intervals without doing any special exercises, and see which one starts walking first. We call this other infant the **control**. Suppose this one starts walking aged 11.50 months (that is, 11 months and 2 weeks). Now, have we answered the question?

It is clear that we’re still not done, because children start walking at all different ages. It could be that we happened to pick a slow child for the exercises, and a particularly fast-developing child for the control. How can we resolve this?

Obviously, the first thing we need to do is to understand how much variability there is among the age at first walking, without imposing an exercise regime. For that, there is no alternative to looking at multiple infants. Here several questions must be considered:

- How many?
- How do we summarise the results of multiple measurements?
- How do we answer the original question?: Do the special exercises make the children learn to walk sooner?

In the original study, the authors had six infants in the **treatment** group (the formal name for the ones who received the exercise — also called the **experimental** group), and six in the control group. (In fact, they had a second control group, that was subject to an alternative exercise regime. But that’s a complication for a later date.) The results are tabulated in Table 1.1. We see that most of the treatment children did start walking earlier

than most of the control children. But not all. The slowest child from the treatment group in fact started walking later than four of the six control children. Should we still be convinced that the treatment is effective? If not, how many more subjects do we need before we can be confident? How would we decide?

Treatment	9.00	9.50	9.75	10.00	13.00	9.50
Control	11.50	12.00	9.00	11.50	13.25	13.00

Table 1.1: Age (in months) at which infants were first able to walk independently. Data from [ZZK72].

The answer is, we can't know for sure. The results are consistent with believing that the treatment had an effect, but they are also consistent with believing that we happened to get a particularly slow group of treatment children, or a fast group of control children, purely by chance. What we need now is a formal way of looking at these results, to tell us how to decide how to draw conclusions from data — “The exercise helped children walk sooner” — and how properly to estimate the confidence we should have in our conclusions — How likely is it that we might have seen a similar result purely by chance, if the exercise did not help? We will use graphical tools, mathematical tools, and logical tools.

## 1.2 Variables

The datasets that Psychologists and Human Scientists collect will usually consist of one or more observations on one or more “variables”.

A **variable** is a property of an object or event that can take on different values.

For example, suppose we collect a dataset by measuring the hair colour, resting heart rate and score on an IQ test of every student in a class. The variables in this dataset would then simply be hair colour, resting heart rate and score on an IQ test, i.e. the variables are the properties that we measured/observed.

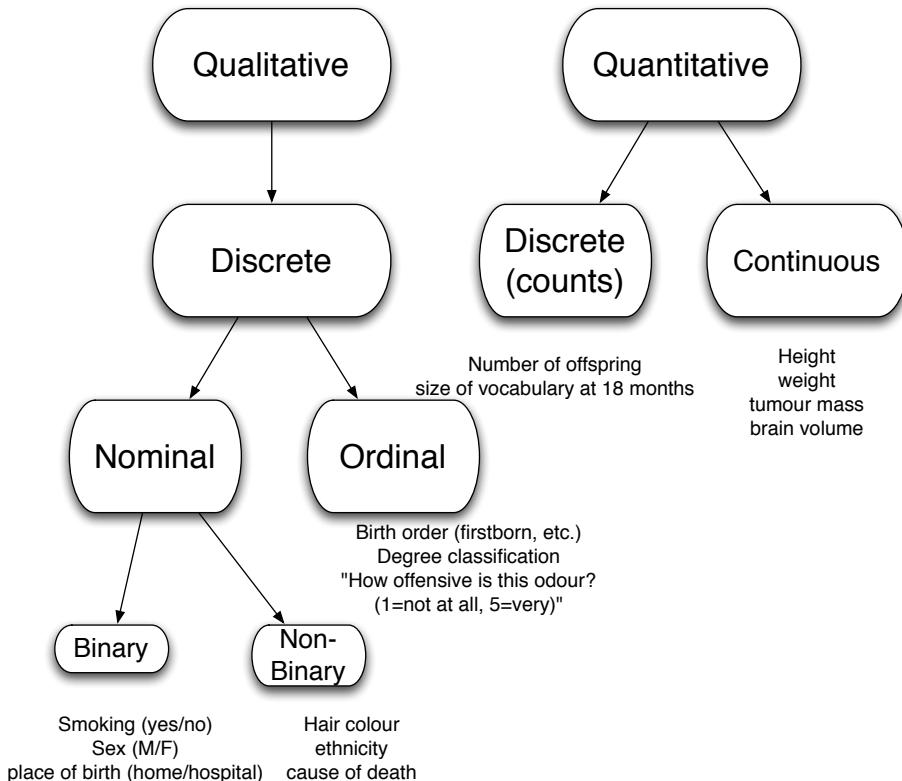


Figure 1.1: A summary of the different data types with some examples.

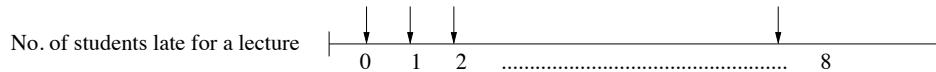
### 1.2.1 Types of variables

There are 2 main types of data/variable (see Figure 1.1)

- **Measurement / Quantitative Data** occur when we measure objects/events to obtain some number that reflects the quantitative trait of interest e.g. when we measure someone's height or weight.
- **Categorical / Qualitative Data** occur when we assign objects into labelled groups or categories e.g. when we group people according to hair colour or race. Often the categories have a natural ordering. For example, in a survey we might group people depending upon whether they agree / neither agree or disagree / disagree with a statement. We call such ordered variables **Ordinal variables**. When the categories are unordered, e.g. hair colour, we have a **Nominal variable**.

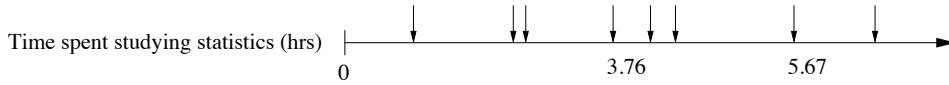
It is also useful to make the distinction between **Discrete** and **Continuous** variables (see Figure 1.2). Discrete variables, such as number of children in a family, or number of peas in a pod, can take on only a limited set of values. (Categorical variables are, of course, always discrete.) Continuous variables, such as height and weight, can take on (in principle) an unlimited set of values.

### Discrete Data



No. of students late for a lecture  
 There are only a limited set of distinct values/categories  
 i.e. we can't have exactly 2.23 students late, only integer values  
 are allowed.

### Continuous Data



Time spent studying statistics (hrs)  
 In theory there are an unlimited set of possible values!  
 There are no discrete jumps between possible values.

Figure 1.2: Examples of Discrete and Continuous variables.

### 1.2.2 Ambiguous data types

The distinctions between the data types described in section 1.2.1 is not always clear-cut. Sometimes, the type isn't inherent to the data, but depends on how you choose to look at it. Consider the experiment described in section 1.1. Think about how the results may have been recorded in the lab notebooks. For each child, it was recorded which group (treatment or control) the child was assigned to, which is clearly a (binary) categorical variable. Then, you might find an entry for each week, recorded the result of the walking test: yes or no, the child could or couldn't walk that week. In principle, this is a long sequence of categorical variables. However, it would

be wise to notice that this sequence consists of a long sequence of “no” followed by a single “yes”. No information is lost, then, if we simply look at the length of the sequence of noes, which is now a *quantitative* variable. Is it discrete or continuous? In principle, the age at which a child starts walking is a continuous variable: there is no fixed set of times at which this can occur. But the variable we have is not the *exact* time of first walking, but the week of the follow-up visit at which the experimenters found that the child could walk, reported in the shorthand that treats 1 week as being 1/4 month. In fact, then, the possible outcomes are a discrete set: 8.0, 8.25, 8.5, .... What this points up, though, is simply that there is no sharp distinction between continuous and discrete. What “continuous” means, in practice, is usually just that there are a large number of possible outcomes. The methods for analysing discrete variables aren’t really distinct from those for continuous variables. Rather, we may have special approaches to analysing a binary variable, or one with a handful of possible outcomes. As the number of outcomes increases, the benefits of considering the “discreteness” disappear, and the methods shade off into the continuous methods.

One important distinction, though, is between categorical and quantitative data. It is obvious that if you have listed each subject’s hair colour, that that needs to be analysed differently from their blood pressure. Where it gets confusing is the case of ordinal categories. For example, suppose we are studying the effect of family income on academic achievement, as measured in degree classification. The possibilities are: first, upper second, lower second, third, pass, and fail. It is clear that they are ordered, so that we want our analysis to take account of the fact that a third is between a fail and a first. The designation even suggests assigning numbers: 1,2,2.5,3,4,5, and this might be a useful shorthand for recording the results. But once we have these numbers, it is tempting to do with them the things we do with numbers: add them up, compute averages, and so on. Keep in mind, though, that we could have assigned any other numbers as well, as long as they have the same order. Do we want our conclusions to depend on the implication that a third is midway between a first and a fail? Probably not.

Suppose you have asked subjects to sniff different substances, and rate them 0, 1, or 2, corresponding to unpleasant, neutral, or pleasant. It’s clear that this is the natural ordering — neutral is between unpleasant and pleasant. The problem comes when you look at the numbers and are tempted to do arithmetic with them. If we had asked subjects how many living grandmothers they have, the answers could be added up to get the total number of grandmothers, which is at least a meaningful quantity. Does the total “pleasant-unpleasant smell” score mean anything? What about the

average score? Is neutral mid-way between unpleasant and pleasant? If half the subjects find it pleasant and half unpleasant, do they have “on average” a neutral response? The answers to these questions are not obvious, and require some careful consideration in each specific instance. Totalling and averaging of arbitrary numbers attached to ordinal categories is a common practice, often carried out heedlessly. It should be done only with caution.

### 1.3 Plotting Data

One of the most important stages in a statistical analysis can be simply to look at your data right at the start. By doing so you will be able to spot characteristic features, trends and outlying observations that enable you to carry out an appropriate statistical analysis. Also, it is a good idea to look at the results of your analysis using a plot. This can help identify if you did something that wasn’t a good idea!

DANGER!! It is easy to become complacent and analyse your data without looking at it. This is a dangerous (and potentially embarrassing) habit to get into and can lead to false conclusions on a given study. The value of plotting your data cannot be stressed enough.

Given that we accept the importance of plotting a dataset we now need the tools to do the job. There are several methods that can be used which we will illustrate with the help of the following dataset.

#### The baby-boom dataset

Forty-four babies (a new record) were born in one 24-hour period at the Mater Mothers’ Hospital in Brisbane, Queensland, Australia, on December 18, 1997. For each of the 44 babies, The Sunday Mail recorded the time of birth, the sex of the child, and the birth weight in grams. The data are shown in Table 1.2, and will be referred to as the “Baby boom dataset.”

While we did not collect this dataset based on a specific hypothesis, if we wished we could use it to answer several questions of interest. For example,

- Do girls weigh more than boys at birth?
- What is the distribution of the number of births per hour?
- Is birth weight related to the time of birth?

Time (min)	Sex	Weight (g)	Time (min)	Sex	Weight (g)
5	F	3837	847	F	3480
64	F	3334	873	F	3116
78	M	3554	886	F	3428
115	M	3838	914	M	3783
177	M	3625	991	M	3345
245	F	2208	1017	M	3034
247	F	1745	1062	F	2184
262	M	2846	1087	M	3300
271	M	3166	1105	F	2383
428	M	3520	1134	M	3428
455	M	3380	1149	M	4162
492	M	3294	1187	M	3630
494	F	2576	1189	M	3406
549	F	3208	1191	M	3402
635	M	3521	1210	F	3500
649	F	3746	1237	M	3736
653	F	3523	1251	M	3370
693	M	2902	1264	M	2121
729	M	2635	1283	M	3150
776	M	3920	1337	F	3866
785	M	3690	1407	F	3542
846	F	3430	1435	F	3278

Table 1.2: The Baby-boom dataset

- Is gender related to the time of birth?
- Are these observations consistent with boys and girls being equally likely?

These are all questions that you will be able to test formally by the end of this course. First though we can plot the data to view what the data might be telling us about these questions.

### 1.3.1 Bar Charts

A Bar Chart is a useful method of summarising Categorical Data. We represent the counts/frequencies/percentages in each category by a bar. Figure 1.3 is a bar chart of gender for the baby-boom dataset. Notice that the bar chart has its axes clearly labelled.



Figure 1.3: A Bar Chart showing the gender distribution in the Baby-boom dataset.

### 1.3.2 Histograms

An analogy

‘A Bar Chart is to Categorical Data as a Histogram is to Measurement Data’

A histogram shows us the “distribution” of the numbers along some scale. A histogram is constructed in the following way

- Divide the measurements into intervals (sometimes called “bins”);
- Determine the number of measurements within each category.
- Draw a bar for each category whose heights represent the counts in each category.

The ‘art’ in constructing a histogram is how to choose the number of bins and the boundary points of the bins. For “small” datasets, it is often feasible to simply look at the values and decide upon sensible boundary points.

For the baby-boom dataset we can draw a histogram of the birth weights (Figure 1.4). To draw the histogram I found the smallest and largest values

$$\text{smallest} = 1745 \quad \text{largest} = 4162$$

There are only 44 weights so it seems reasonable to take 6 bins

1500-2000   2000-2500   2500-3000   3000-3500   3500-4000   4000-4500

Using these categories works well, the histogram shows us the shape of the distribution and we notice that distribution has an extended left ‘tail’.

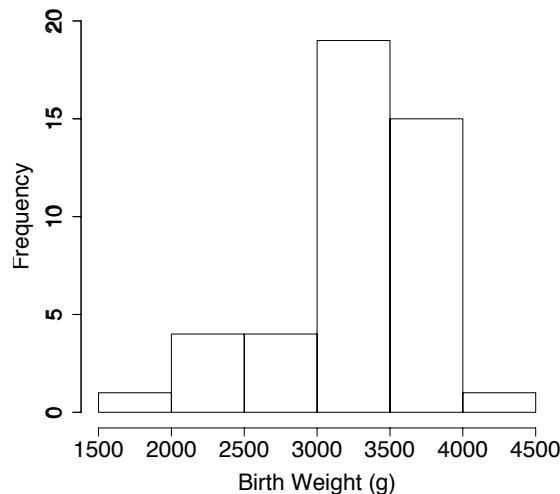


Figure 1.4: A Histogram showing the birth weight distribution in the Baby-boom dataset.

Too few categories and the details are lost. Too many categories and the overall shape is obscured by haphazard details (see Figure 1.5).

In Figure 1.6 we show some examples of the different shapes that histograms can take. One can learn quite a lot about a set of data by looking just at the shape of the histogram. For example, Figure 1.6(c) shows the percentage of the tuberculosis drug isoniazid that is acetylated in the livers of 323 patients after 8 hours. Unacetylated isoniazid remains longer in the blood, and can contribute to toxic side effects. It is interesting, then, to

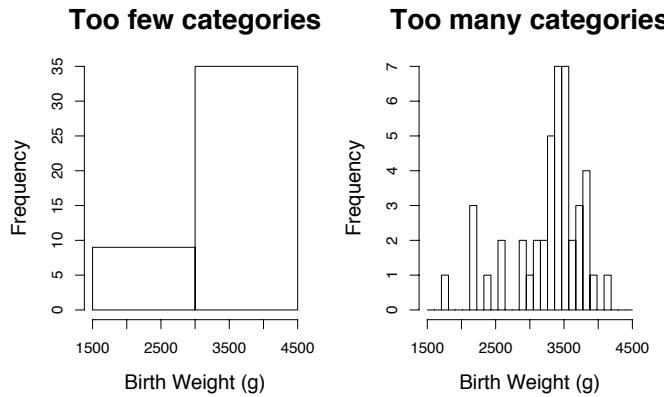


Figure 1.5: Histograms with too few and too many categories respectively.

notice that there is a wide range of rates of acetylation, from patients who acetylate almost all, to those who acetylate barely one fourth of the drug in 8 hours. Note that there are two peaks — this kind of distribution is called **bimodal** — which points to the fact that there is a subpopulation who lacks a functioning copy of the relevant gene for efficiently carrying through this reaction.

So far, we have taken the bins to all have the same width. Sometimes we might choose to have unequal bins, and more often we may be forced to have unequal bins by the way the data are delivered. For instance, suppose we did not have the full table of data, but were only presented with the following table: What is the best way to make a histogram from these data?

Bin	1500–2500g	2500–3000g	3000–3500g	3500–4500g
Number of Births	5	4	19	16

Table 1.3: Baby-boom weight data, allocated to unequal bins.

We could just plot rectangles whose heights are the frequencies. We then end up with the picture in Figure 1.7(a). Notice that the shape has changed substantially, owing to the large boxes that correspond to the widened bins. In order to preserve the shape — which is the main goal of a histogram — we want the area of a box to correspond to the contents of the bin, rather than the height. Of course, this is the same when the bin widths are equal. Otherwise, we need to switch from the **frequency scale** to **density scale**,

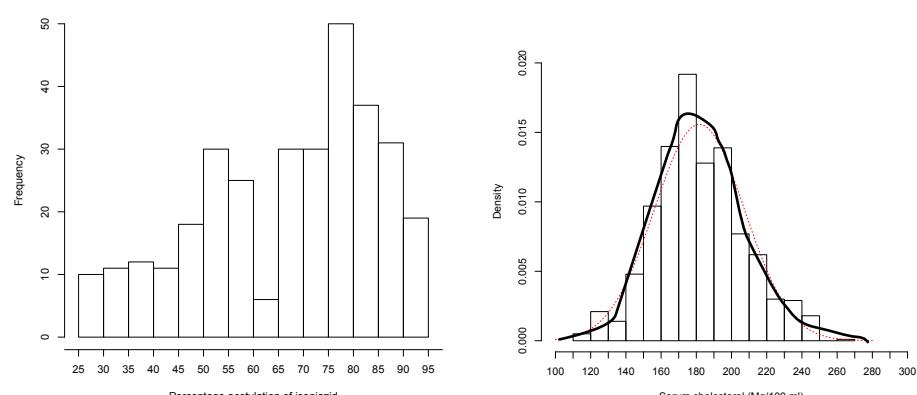
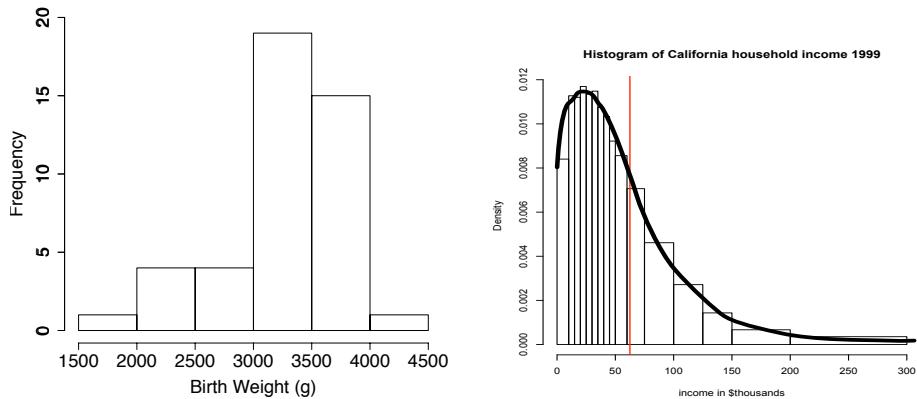


Figure 1.6: Examples of different shapes of histograms.

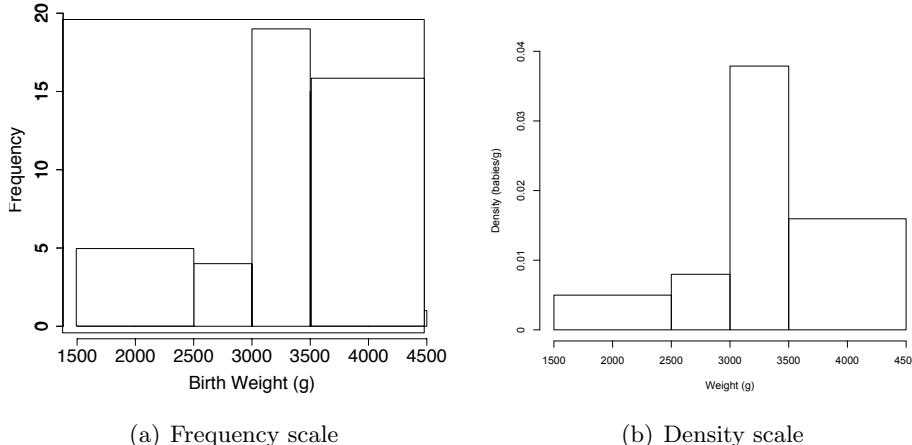


Figure 1.7: Same data plotted in frequency scale and density scale. Note that the density scale histogram has the same shape as the plot from the data with equal bin widths.

in which the height of a box is not the number of observations in the bin, but the number of observations per unit of measurement. This gives us the picture in Figure 1.7(b), which has a very similar shape to the histogram with equal bin-widths.

Thus, for the data in Table 1.3 we would calculate the height of the first rectangle as

$$\text{Density} = \frac{\text{Number of births}}{\text{width of bin}} = \frac{5 \text{ babies}}{1000g} = 0.005.$$

The complete computations are given in Table 1.4, and the resulting histogram is in Figure 1.7(b).

Bin	1500–2500g	2500–3000g	3000–3500g	3500–4500g
Number of Births	5	4	19	16
Density	0.005	0.008	0.038	0.016

Table 1.4: Computing a histogram in density scale

A histogram in density scale is constructed in the following way

- Divide the measurements into bins (unless these are already given);

- Determine the number of measurements within each category (Note: the “number” can also be a percentage. Often the exact numbers are unavailable, but you can simply act as though there were 100 observations);
- For each bin, compute the *density*, which is simply the number of observations divided by the width of a bin;
- Draw a bar for each bin whose height represent the density in each bin. The area of the bar will correspond to the number of observations in the bin.

### 1.3.3 Cumulative and Relative Cumulative Frequency Plots and Curves

A **cumulative frequency plot** is very similar to a histogram. In a cumulative frequency plot the height of the bar in each interval represents the total count of observations within interval and *lower than* the interval (see Figure 1.8)

In a **cumulative frequency curve** the cumulative frequencies are plotted as points at the upper boundaries of each interval. It is usual to join up the points with straight lines (see Figure 1.8).

Relative cumulative frequencies are simply cumulative frequencies divided by the total number of observations (so relative cumulative frequencies always lie between 0 and 1). Thus **relative cumulative frequency plots and curves** just use relative cumulative frequencies rather than cumulative frequencies. Such plots are useful when we wish to compare two or more distributions on the same scale.

Consider the histogram of birth weight shown in Figure 1.4. The frequencies, cumulative frequencies and relative cumulative frequencies of the intervals are given in Table 1.5.

### 1.3.4 Dot plot

A Dot Plot is a simple and quick way of visualising a dataset. This type of plot is especially useful if data occur in groups and you wish to quickly visualise the differences between the groups. For example, Figure 1.9 shows

Interval	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000-4500
Frequency	1	4	4	19	15	1
Cumulative Frequency	1	5	9	28	43	44
Relative Cumulative Frequency	0.023	0.114	0.205	0.636	0.977	1.0

Table 1.5: Frequencies and Cumulative frequencies for the histogram in Figure 1.4.

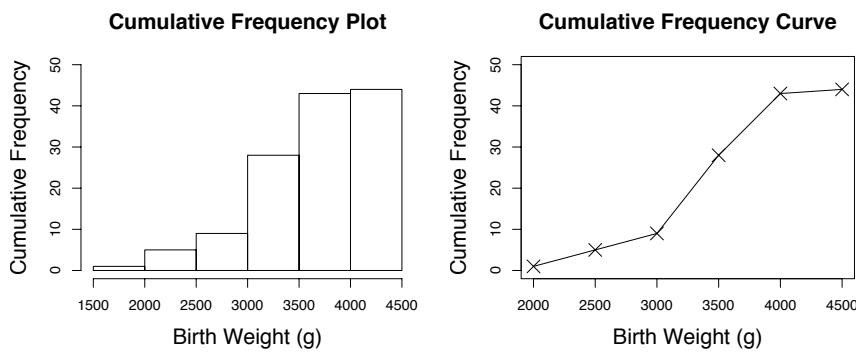


Figure 1.8: Cumulative frequency curve and plot of birth weights for the baby-boom dataset.

a dot plot of birth weights grouped by gender for the baby-boom dataset. The plot suggests that girls may be lighter than boys at birth.

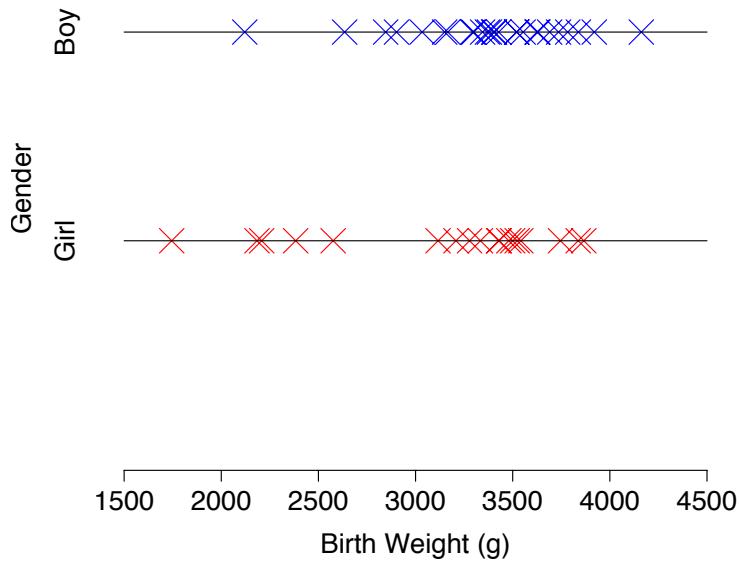


Figure 1.9: A Dot Plot showing the birth weights grouped by gender for the baby-boom dataset.

### 1.3.5 Scatter Plots

Scatter plots are useful when we wish to visualise the relationship between two measurement variables.

To draw a scatter plot we

- Assign one variable to each axis.
- Plot one point for each pair of measurements.

For example, we can draw a scatter plot to examine the relationship between birth weight and time of birth (Figure 1.10). The plot suggests that there is little relationship between birth weight and time of birth.

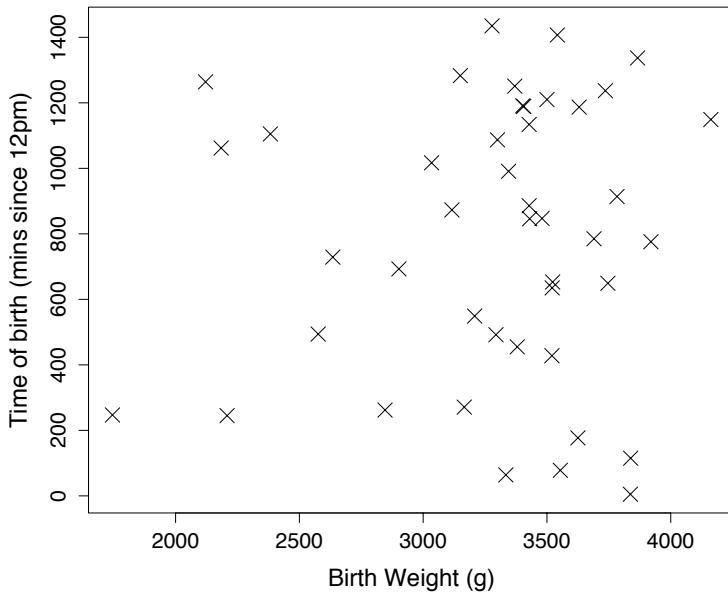


Figure 1.10: A Scatter Plot of birth weights versus time of birth for the baby-boom dataset.

### 1.3.6 Box Plots

Box Plots are probably the most sophisticated type of plot we will consider. To draw a Box Plot we need to know how to calculate certain “summary measures” of the dataset covered in the next section. We return to discuss Box Plots in Section 1.5.

## 1.4 Summary Measures

In the previous section we saw how to use various graphical displays in order to explore the structure of a given dataset. From such plots we were able to observe the general shape of the “distribution” of a given dataset and compare visually the shape of two or more datasets.

Consider the histogram in Figure 1.11. Comparing the first and second histograms we see that the distributions have the same shape or spread but that the center of the distribution is different. Roughly, by eye, the centers differ in value by about 10. Comparing first and third histograms we see

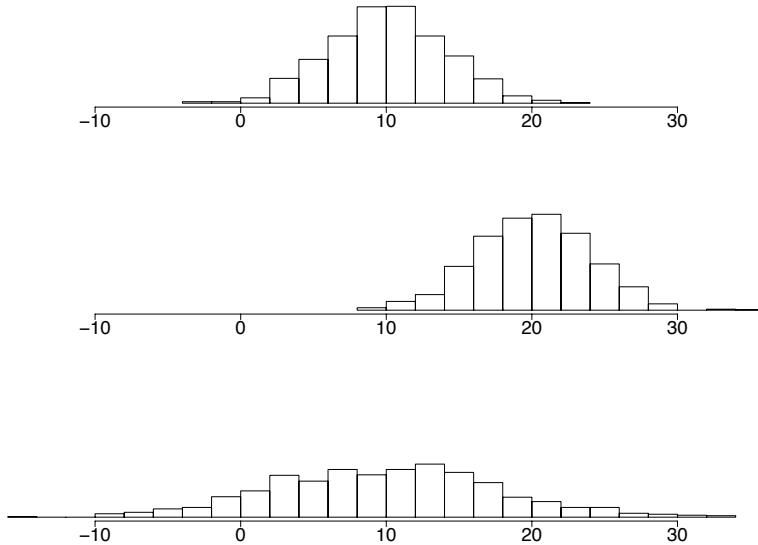


Figure 1.11: Comparing shapes of histograms

that the distributions seem to have roughly the same center but that the data plotted in the third are more spread out than in the first. Obviously, comparing second and the third we observe differences in both the center and the spread of the distribution.

While it is straightforward to compare two distributions “by eye”, placing the two histograms next to each other, it is clear that this would be difficult to do with ten or a hundred different distributions. For example, Figure 1.6(a) shows a histogram of 1999 incomes in California. Suppose we wanted to compare incomes among the 50 US states, or see how incomes developed annually from 1980 to 2000, or compare these data to incomes in 20 other industrialised countries. Laying out the histograms and comparing them would not be very practical.

Instead, we would like to have single numbers that measure

- the ‘center’ point of the data.
- the ‘spread’ of the data.

These two characteristics of a set of data (the center and spread) are the simplest measures of its shape. Once calculated we can make precise statements about how the centers or spreads of two datasets differ. Later we will

learn how to go a stage further and ‘test’ whether two variables have the same center point.

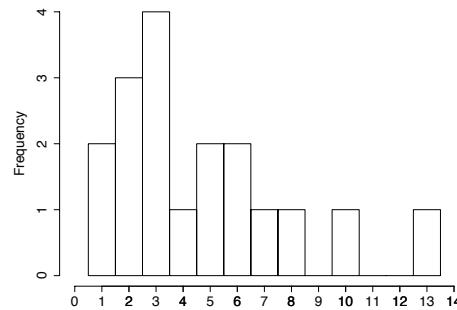
#### 1.4.1 Measures of location (Measuring the center point)

There are 3 main measures of the center of a given set of (measurement) data

- **The Mode** of a set of numbers is simply the most common value e.g. the mode of the following set of numbers

1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 5, 5, 6, 6, 7, 8, 10, 13

is 3. If we plot a histogram of this data



we see that the mode is the peak of the distribution and is a reasonable representation of the center of the data. If we wish to calculate the mode of continuous data one strategy is to group the data into adjacent intervals and choose the modal interval i.e. draw a histogram and take the modal peak. This method is sensitive to the choice of intervals and so care should be taken so that the histogram provides a good representation of the shape of the distribution.

The Mode has the advantage that it is always a score that actually occurred and can be applied to nominal data, properties not shared by the median and mean. A disadvantage of the mode is that there may two or more values that share the largest frequency. In the case of two modes we would report both and refer to the distribution as *bimodal*.

- **The Median** can be thought of as the ‘middle’ value i.e. the value for which 50% of the data fall below when arranged in numerical order. For example, consider the numbers

$$15, 3, 9, 21, 1, 8, 4,$$

When arranged in numerical order

$$1, 3, 4, \boxed{8}, 9, 15, 21$$

we see that the median value is 8. If there were an even number of scores e.g.

$$1, 3, \boxed{4, 8}, 9, 15$$

then we take the midpoint of the two middle values. In this case the median is  $(4+8)/2 = 6$ . In general, if we have  $N$  data points then the **median location** is defined as follows:

$$\text{Median Location} = \frac{(N+1)}{2}$$

For example, the median location of 7 numbers is  $(7+1)/2 = 4$  and the median of 8 numbers is  $(8+1)/2 = 4.5$  i.e. between observation 4 and 5 (when the numbers are arranged in order).

A major advantage of the median is the fact that it is unaffected by extreme scores (a point it shares with the mode). We say the median is **resistant** to outliers. For example, the median of the numbers

$$1, 3, 4, \boxed{8}, 9, 15, 99999$$

is still 8. This property is very useful in practice as *outlying* observations can and do occur (Data is messy remember!).

- **The Mean** of a set of scores is the sum<sup>1</sup> of the scores divided by the number of scores. For example, the mean of

$$1, 3, 4, 8, 9, 15 \quad \text{is} \quad \frac{1+3+4+8+9+15}{6} = 6.667 \quad (\text{to 3 dp})$$

In mathematical notation, the mean of a set of  $n$  numbers  $x_1, \dots, x_n$  is denoted by  $\bar{x}$  where

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{or} \quad \bar{x} = \frac{\sum x}{n} \quad (\text{in the formula book})$$

---

<sup>1</sup>The total when we add them all up

See the appendix for a brief description of the summation notation ( $\sum$ )

---

The mean is the most widely used measure of location. Historically, this is because statisticians can write down equations for the mean and derive nice theoretical properties for the mean, which are much harder for the mode and median. A disadvantage of the mean is that it is not resistant to outlying observations. For example, the mean of

$$1, 3, 4, 8, 7, 15, 99999$$

is 14323.57, whereas the median (from above) is 8.

Sometimes discrete measurement data are presented in the form of a frequency table in which the frequencies of each value are given. Remember, the mean is the sum of the data divided by the number of observations. To calculate the sum of the data we simply multiply each value by its frequency and sum. The number of observations is calculated by summing the frequencies.

For example, consider the following frequency table

Data (x)	1	2	3	4	5	6
Frequency (f)	2	4	6	7	4	1

Table 1.6: A frequency table.

We calculate the sum of the data as

$$(2 \times 1) + (4 \times 2) + (6 \times 3) + (7 \times 4) + (4 \times 5) + (1 \times 6) = 82$$

and the number of observations as

$$2 + 4 + 6 + 7 + 4 + 1 = 24$$

The the mean is given by

$$\bar{x} = \frac{82}{24} = 3.42 \quad (2 \text{ dp})$$

---

In mathematical notation the formula for the mean of frequency data is given by

$$\bar{x} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i} \quad \text{or} \quad \bar{x} = \frac{\sum f x}{\sum f}$$


---

### The relationship between the mean, median and mode

In general, these three measures of location will differ but for certain datasets with characteristic ‘shapes’ we will observe simple patterns between the three measures (see Figure 1.12).

- If the distribution of the data is **symmetric** then the mean, median and mode will be very close to each other.

$$\text{MODE} \approx \text{MEDIAN} \approx \text{MEAN}$$

- If the distribution is **positively skewed** or **right skewed** i.e. the data has an extended right tail, then

$$\text{MODE} < \text{MEDIAN} < \text{MEAN}.$$

Income data, as in Figure 1.6(b), tends to be right-skewed. The mean is shown by a red

- If the distribution is **negatively skewed** or **left skewed** i.e. the data has an extended left tail, as then

$$\text{MEAN} < \text{MEDIAN} < \text{MODE}.$$

- If the

### The mid-range

There is actually a fourth measure of location that can be used (but rarely is). The **Mid-Range** of a set of data is half way between the smallest and largest observation i.e. half the **range** of the data. For example, the mid-range of

$$1, 3, 4, 8, 9, 15, 21$$

is  $(1 + 21) / 2 = 11$ . The mid-range is rarely used because it is not resistant to outliers and by using only 2 observations from the dataset it takes no account of how spread out most of the data are.

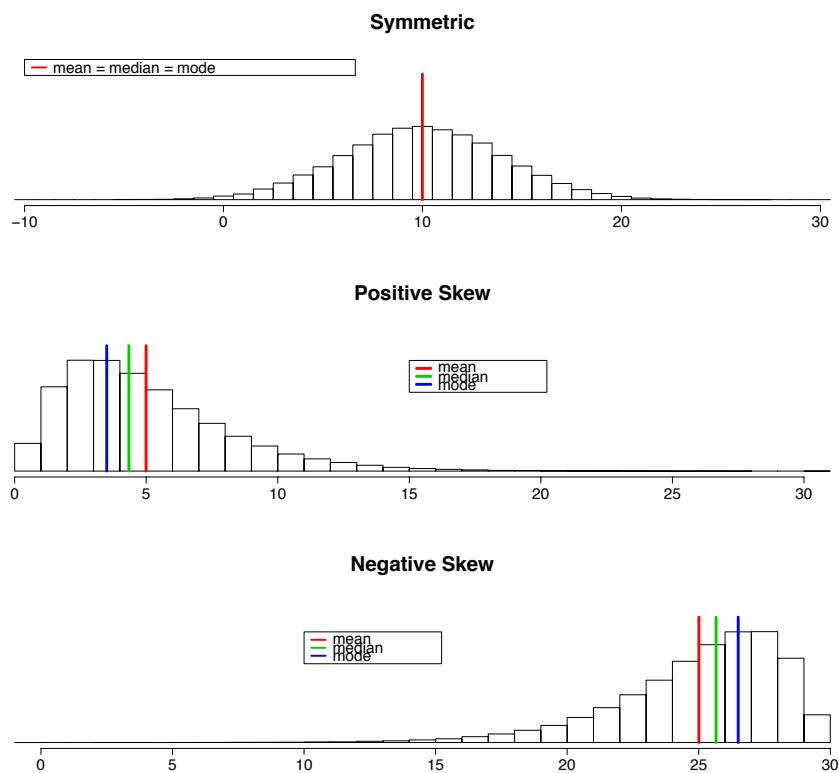


Figure 1.12: The relationship between the mean, median and mode.

### 1.4.2 Measures of dispersion (Measuring the spread)

#### The Interquartile Range (IQR) and Semi-Interquartile Range (SIQR)

The Interquartile Range (IQR) of a set of numbers is defined to be the range of the middle 50% of the data (see Figure 1.13). The Semi-Interquartile Range (SIQR) is simply half the IQR.

---

We calculate the IQR in the following way:

- Calculate the 25% point (**1st quartile**) of the dataset. The location of the 1st quartile is defined to be the  $(\frac{N+1}{4})$ th data point.
- Calculate the 75% point (**3rd quartile**) of the dataset. The location of the 3rd quartile is defined to be the  $(\frac{3(N+1)}{4})$ th data point<sup>2</sup>.
- Calculate the IQR as

$$\text{IQR} = \text{3rd quartile} - \text{1st quartile}$$


---

**Example 1** Consider the set of 11 numbers (which have been arranged in order)

$$10, 15, 18, 33, 34, 36, 51, 73, 80, 86, 92.$$

The 1st quartile is the  $\frac{(11+1)}{4} = 3$ rd data point = 18

The 3rd quartile is the  $\frac{3(11+1)}{4} = 9$ th data point = 80

$$\begin{aligned}\Rightarrow \text{IQR} &= 80 - 18 = 62 \\ \Rightarrow \text{SIQR} &= 62 / 2 = 31.\end{aligned}$$

What do we do when the number of points +1 isn't divisible by 4? We interpolate, just like with the median. Suppose we take off the last data point, so the list of data becomes

$$10, 15, 18, 33, 34, 36, 51, 73, 80, 86.$$

What is the 1st quartile? We're now looking for the  $\frac{(10+1)}{4} = 2.75$  data point. This should be 3/4 of the way from the 2nd data point to the 3rd. The distance from 15 to 18 is 3. 1/4 of the way is .75, and 3/4 of the way is 2.25. So the 1st quartile is 17.25.

---

<sup>2</sup>The **2nd quartile** is the 50% point of the dataset i.e. the median.

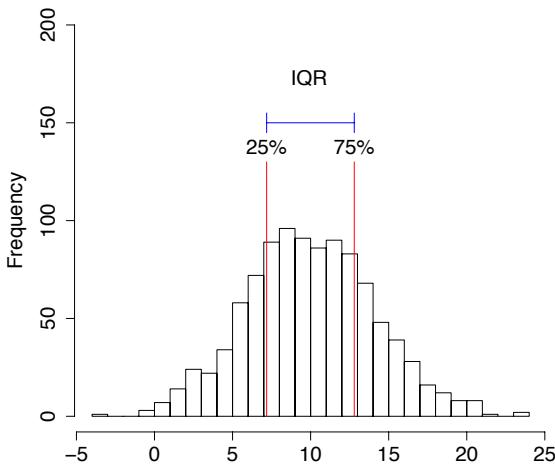


Figure 1.13: The Interquartile Range.

### The Mean Deviation

To measure the spread of a dataset it seems sensible to use the ‘deviation’ of each data point from the mean of the distribution (see Figure 1.14). The deviation of each data point from the mean is simply the data point minus the mean.

For example, for deviations of the set of numbers

$$10, 15, 18, 33, 34, 36, 51, 73, 80, 86, 92$$

which have mean 48 are given in Table 1.7.

The Mean Deviation of a set of numbers is simply mean of deviations.

In mathematical notation this is written as

$$\frac{\sum_{i=1}^n (x_i - \bar{x})}{n}$$

At first this sounds like a good way of assessing the spread since you might think that large spread gives rise to larger deviations and thus a larger

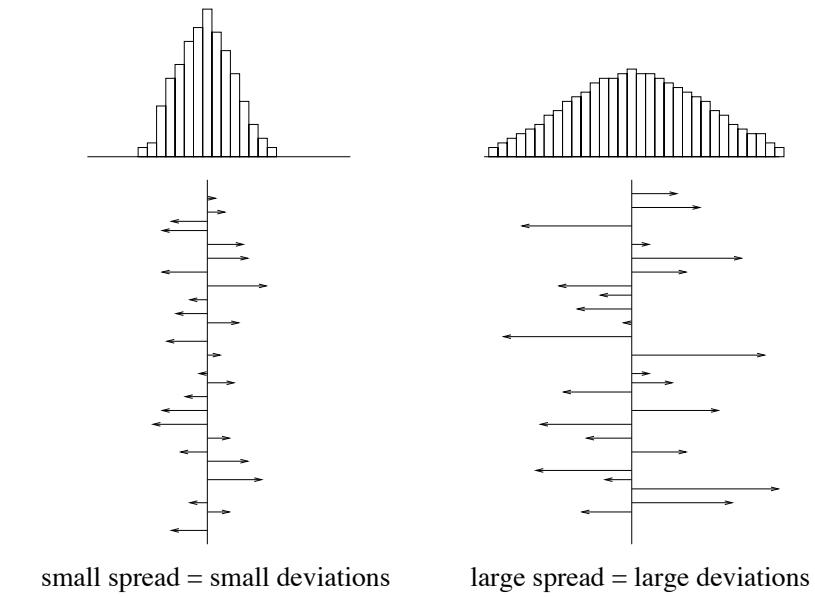


Figure 1.14: The relationship between spread and deviations..

mean deviation. In fact, though, the mean deviation is *always* zero. The positive and negative deviations cancel each other out exactly. Even so, the deviations still contain useful information about the spread, we just have to find a way of using the deviations in a sensible way.

### Mean Absolute Deviation (MAD)

We solve the problem of the deviations summing to zero by considering the *absolute values* of the deviations. The absolute value of a number is the value of that number with any minus sign removed, e.g.  $|-3| = 3$ . We then measure spread using the mean of the absolute deviations, denoted (MAD).

This can be written in mathematical notation as

$$\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} \quad \text{or} \quad \frac{\sum |x - \bar{x}|}{n}$$

**Note** The second formula is just a short hand version of the first (See the Appendix).

We calculate the MAD in the following way (see Table 1.7 for an example)

Data $x$	Deviations $x - \bar{x}$	Absolute Deviations $ x - \bar{x} $	Squared Deviations $(x - \bar{x})^2$
10	$10 - 48 = -38$	38	1444
15	$15 - 48 = -33$	33	1089
18	$18 - 48 = -30$	30	900
33	$33 - 48 = -15$	15	225
34	$34 - 48 = -14$	14	196
36	$36 - 48 = -12$	12	144
51	$51 - 48 = 3$	3	9
73	$73 - 48 = 25$	25	625
80	$80 - 48 = 32$	32	1024
86	$86 - 48 = 38$	38	1444
92	$92 - 48 = 44$	44	1936
Sum = 528	Sum = 0	Sum = 284	Sum = 9036
$\sum x = 528$	$\sum(x - \bar{x}) = 0$	$\sum  x - \bar{x}  = 284$	$\sum(x - \bar{x})^2 = 9036$

Table 1.7: Deviations, Absolute Deviations and Squared Deviations.

- Calculate the mean of the data,  $\bar{x}$
- Calculate the deviations by subtracting the mean from each value,  $x - \bar{x}$
- Calculate the absolute deviations by removing any minus signs from the deviations,  $|x - \bar{x}|$ .
- Sum the absolute deviations,  $\sum |x - \bar{x}|$ .
- Calculate the MAD by dividing the sum of the absolute deviations by the number of data points,  $\sum |x - \bar{x}|/n$ .

---

From Table 1.7 we see that the sum of the absolute deviations of the numbers in Example 1 is 284 so

$$\text{MAD} = \frac{284}{11} = 25.818 \quad (\text{to 3dp})$$

### The Sample Variance ( $s^2$ ) and Population Variance ( $\sigma^2$ )

Another way to ensure the deviations don't sum to zero is to look at the *squared* deviations i.e. the square of a number is always positive. Thus

another way of measuring the spread is to consider the mean of the squared deviations, called the *variance*

---

If our dataset consists of the whole population (a rare occurrence) then we can calculate the population variance  $\sigma^2$  (said ‘sigma squared’) as

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad \text{or} \quad \sigma^2 = \frac{\sum(x - \bar{x})^2}{n}$$

When we just have a sample from the population (most of the time) we can calculate the sample variance  $s^2$  as

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad \text{or} \quad s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

**Note** We divide by  $n - 1$  when calculating the sample variance as then  $s^2$  is a ‘better estimate’ of the population variance  $\sigma^2$  than if we had divided by  $n$ . We will see why later.

For frequency data (see Table 1.6) the formula is given by

$$s^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i - 1} \quad \text{or} \quad s^2 = \frac{\sum f(x - \bar{x})^2}{\sum f - 1}$$

---

We can calculate  $s^2$  in the following way (see Table 1.7)

- Calculate the deviations by subtracting the mean from each value,  $x - \bar{x}$
- Calculate the squared deviations,  $(x - \bar{x})^2$ .
- Sum the squared deviations,  $\sum(x - \bar{x})^2$ .
- Divide by  $n - 1$ ,  $\sum(x - \bar{x})^2 / (n - 1)$ .

From Table 1.7 we see that the sum of the squared deviations of the numbers in Example 1 is 9036 so

$$s^2 = \frac{9036}{11 - 1} = 903.6$$

### The Sample Standard Deviation ( $s$ ) and Population Standard Deviation ( $\sigma$ )

Notice how the sample variance in Example 1 is much higher than the SIQR and the MAD.

$$\text{SIQR} = 31 \quad \text{MAD} = 25.818 \quad s^2 = 903.6$$

This happens because we have squared the deviations transforming them to a quite different scale. We can recover the scale of the original data by simply taking the square root of the sample (population) variance.

Thus we define the sample standard deviation  $s$  as

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

and we define the population standard deviation  $\sigma$  as

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Returning to Example 1 the sample standard deviation is

$$s = \sqrt{903.6} = 30.05 \quad (\text{to 2dp})$$

which is comparable with the SIQR and the MAD.

## 1.5 Box Plots

As we mentioned earlier a Box Plot (sometimes called a Box-and-Whisker Plot) is a relatively sophisticated plot that summarises the distribution of a given dataset.

A box plot consists of three main parts

- A box that covers the middle 50% of the data. The edges of the box are the 1st and 3rd quartiles. A line is drawn in the box at the median value.
- Whiskers that extend out from the box to indicate how far the data extend either side of the box. The whiskers should extend no further than 1.5 times the length of the box, i.e. the maximum length of a whisker is 1.5 times the IQR.

- All points that lie outside the whiskers are plotted individually as outlying observations.

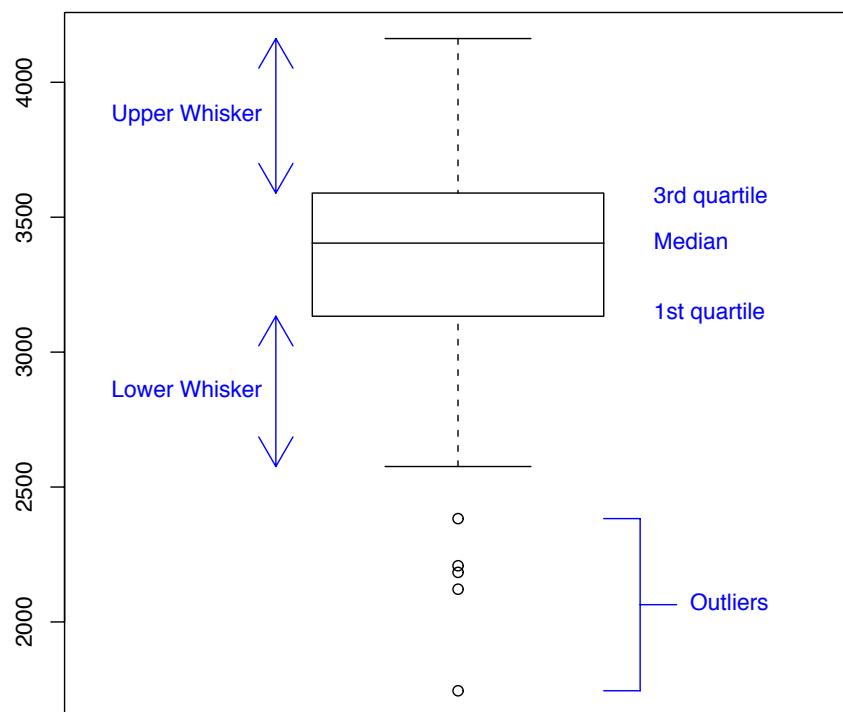


Figure 1.15: A Box Plot of birth weights for the baby-boom dataset showing the main points of plot.

Plotting box plots of measurements in different groups side by side can be illustrative. For example, Figure 1.16 shows box plots of birth weight for each gender side by side and indicates that the distributions have quite different shapes.

Box plots are particularly useful for comparing multiple (but not very many!) distributions. Figure 1.17 shows data from 14 years, of the total

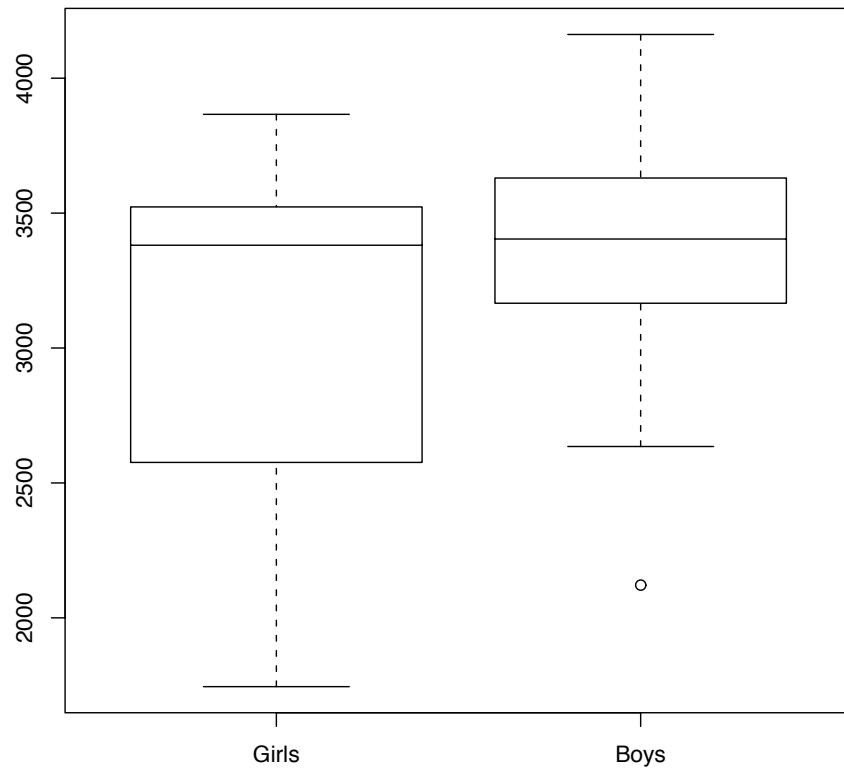


Figure 1.16: A Box Plot of birth weights by gender for the baby-boom dataset.

number of births each day (5113 days in total) in Quebec hospitals. By summarising the data in this way, it becomes clear that there is a substantial difference between the numbers of births on weekends and on weekdays. We see that there is a wide variety of numbers of births, and considerable overlap among the distributions, but the medians for the weekdays are far outside the range of most of the weekend numbers.

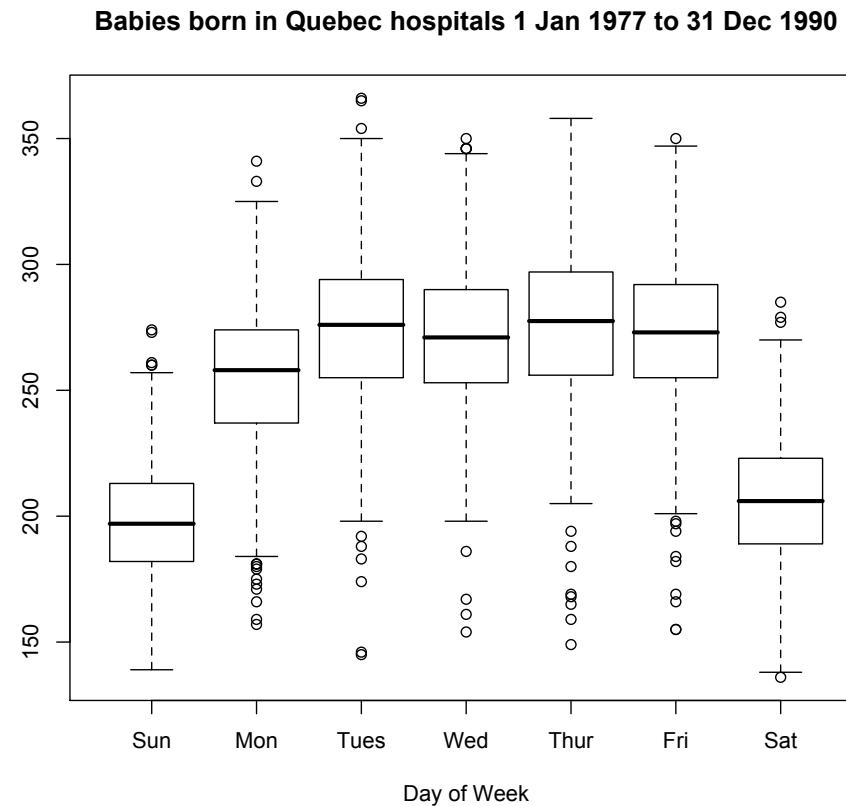


Figure 1.17: Daily numbers of births in Quebec hospitals, 1 Jan 1977 to 31 Dec 1990.

## 1.6 Appendix

### 1.6.1 Mathematical notation for variables and samples

Mathematicians are lazy. They can't be bothered to write everything out in full so they have invented a language/notation in which they can express what they mean in a compact, quick to write down fashion. This is a good thing. We don't have to study maths every day to be able to use a bit of the language and make our lives easier. For example, suppose we are interested in comparing the resting heart rate of 1st year Psychology and Human Sciences students. Rather than keep on referring to variables 'the resting heart rate of 1st year Psychology students' and 'the resting heart rate of 1st year

Human Sciences students' we can simple denote

$$\begin{aligned} X &= \text{the resting heart rate of 1st year Psychology students} \\ Y &= \text{the resting heart rate of 1st year Human Sciences students} \end{aligned}$$

and refer to the variables X and Y instead.

In general, we use capital letters to denote variables. If we have a sample of a variable the we use small letters to denote the sample. For example, if we go and measure the resting heart rate of 1st year Psychology and Human Sciences students in Oxford we could denote the  $p$  measurements on Psychologists as

$$x_1, x_2, x_3, \dots, x_p$$

and the  $h$  measurements on Human Scientists as

$$y_1, y_2, y_3, \dots, y_h$$

### 1.6.2 Summation notation

One of the most common letters in the Mathematicians alphabet is the Greek letter **sigma** ( $\sum$ ), which is used to denote summation. It translates to “add up what follows”. Usually, the limits of the summation are written below and above the symbol. So,

$$\sum_{i=1}^5 x_i \quad \text{reads “add up the } x_i\text{s from } i = 1 \text{ to } i = 5\text{”}$$

or

$$\sum_{i=1}^5 x_i = (x_1 + x_2 + x_3 + x_4 + x_5)$$

If we have some actual data then we know the values of the  $x_i$ s

$$x_1 = 3 \quad x_2 = 6 \quad x_3 = 1 \quad x_4 = 7 \quad x_5 = 6$$

and we can calculate the sum as

$$\sum_{i=1}^5 x_i = (3 + 2 + 1 + 7 + 6) = 19$$

If the limits of the summation are obvious within context the the notation is often abbreviated to

$$\sum x = 19$$



## Lecture 2

# Probability I

In this and the following lecture we will learn about

- why we need to learn about probability
- what probability is
- how to assign probabilities
- how to manipulate probabilities and calculate probabilities of complex events

### 2.1 Why do we need to learn about probability?

In Lecture 1 we discussed why we need to study statistics, and we saw that statistics plays a crucial role in the scientific process (see figure 2.1). We saw that we use a sample from the population in order to test our hypothesis. There will usually be a very large number of possible samples we could have taken and the conclusions of the statistical test we use will depend on the exact sample we take. It might happen that the sample we take leads us to make the wrong conclusion about the population. Thus we need to know what the chances are of this happening. Probability can be thought of as the study of chance.

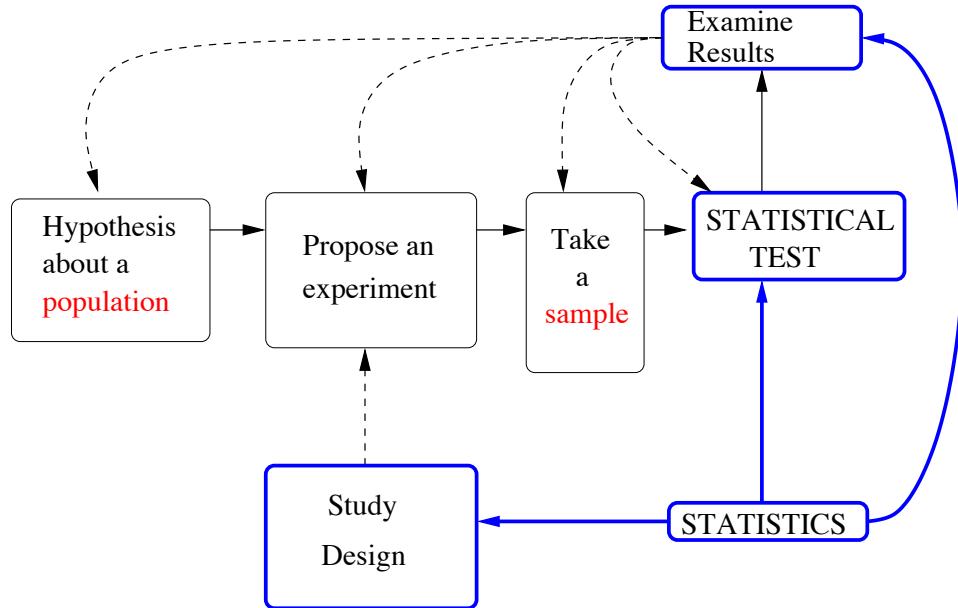


Figure 2.1: The scientific process and role of statistics in this process.

### Example 2.1: Random controlled experiment

The Anturane Reinfarction Trial (ART) was a famous study of a drug treatment (anturane) for the aftereffects of myocardial infarction [MDF<sup>+</sup>81]. Out of 1629 patients, about half (813) were selected to receive anturane; the other 816 patients received an ineffectual (“placebo”) pill. The results are summarised in Table 2.1.

Table 2.1: Results of the Anturane Reinfarction Trial.

	Treatment (anturane)	Control (placebo)
# patients	813	816
deaths	74	89
% mortality	9.1%	10.9%

Imagine the following dialogue:

**Drug Company:** Every hospital needs to use anturane. It saves patients' lives.

**Skeptical Bureaucrat:** The effect looks pretty small: 15 out of about 800 patients. And the drug is pretty expensive.

**DC:** Is money all you bean counters can think about? We reduced mortality by 16%.

**SB:** It was only 2% of the total.

**DC:** We saved 2% of the patients! What if one of them was your mother?

**SB:** I'm all in favour of saving 2% more patients. I'm just wondering: You flipped a coin to decide which patients would get the anturane. What if the coins had come up differently? Might we just as well be here talking about how anturane had killed 2% of the patients?

How can we resolve this argument? 163 patients died. Suppose anturane has no effect. Could the apparent benefit of anturane simply reflect the random way the coins fell? Or would such a series of coin flips have been simply too unlikely to countenance? To answer this question, we need to know how to measure the likelihood (or "probability") of sequences of coin flips.

Imagine a box, with cards in it, each one having written on it one way in which the coin flips could have come out, and the patients allocated to treatments. How many of those coinflip cards would have given us the impression that anturane performed well, purely because many of the patients who died happened to end up in the Control (placebo) group? It turns out that it's more than 20% of the cards, so it's really not very unlikely at all.

To figure this out, we are going to need to understand

1. How to enumerate all the ways the coins could come up. How many ways are there? The number depends on the exact procedure, but if we flip one coin for each patient, the number of cards in the box would be  $2^{1629}$ , which is vastly

more than the number of atoms in the universe. Clearly, we don't want to have to count up the cards individually.

2. How coin flips get associated with a result, as measured in apparent success or failure of anturane. Since the number of "cards" is so large, we need to do this without having to go through the results one by one.

■

### Example 2.2: Baby-boom

Consider the Baby-boom dataset we saw in Lecture 1. Suppose we have a hypothesis that in the population boys weigh more than girls at birth. We can use our sample of boys and girls to examine this hypothesis. One intuitive way of doing this would be to calculate the mean weights of the boys and girls in the sample and compare the difference between these two means

Sample mean of boys weights =  $\bar{x}_{\text{boys}} = 3375$

Sample mean of girls weights =  $\bar{x}_{\text{girls}} = 3132$

$$\Rightarrow D = \bar{x}_{\text{boys}} - \bar{x}_{\text{girls}} = 3375 - 3132 = 243$$

Does this allow us to conclude that in the population boys are born heavier than girls? On what scale to we assess the size of  $D$ ? Maybe boys and girls weigh the same at birth and we obtained a sample with heavier boys just by chance. To be able to conclude confidently that boys in the population are heavier than girls we need to know what the chances are of obtaining a difference between the means that is 243 or greater, i.e. we need to know the probability of getting such a large value of  $D$ . If the chances are small then we can be confident that in the population boys are heavier than girls on average at birth. ■

## 2.2 What is probability?

The examples we have discussed so far look very complicated. They aren't really, but in order to see the simple underlying structure, we need to introduce a few new concepts. To do so, we want to work with a much simpler example:

### Example 2.3: Rolling a die

Consider a simple **experiment** of rolling a fair six-sided die.

When we toss the die there are six possible outcomes i.e. 1, 2, 3, 4, 5 and 6. We say that the **sample space** of our experiment is the set  $S = \{1, 2, 3, 4, 5, 6\}$ .

The outcome "the top face shows a three" is the **sample point** 3.

The **event**  $A_1$ , that the die shows an even number is the subset  $A_1 = \{2, 4, 6\}$  of the sample space.

The **event**  $A_2$  that the die shows a number larger than 4 is the subset  $A_2 = \{5, 6\}$  of  $S$ . ■

#### 2.2.1 Definitions

The example above introduced some terminology that we will use repeatedly when we talk about probabilities.

An **experiment** is some activity with an observable outcome.

The set of all possible outcomes of the experiment is called the **sample space**.

A particular outcome is called a **sample point**.

A collection of possible outcomes is called an **event**.

#### 2.2.2 Calculating simple probabilities

Simply speaking, the probability of an event is a number between 0 and 1, inclusive, that indicates how likely the event is to occur.

In some settings (like the example of the fair die considered above) it is natural to assume that all the sample points are equally likely.

In this case, we can calculate the probability of an event  $A$  as

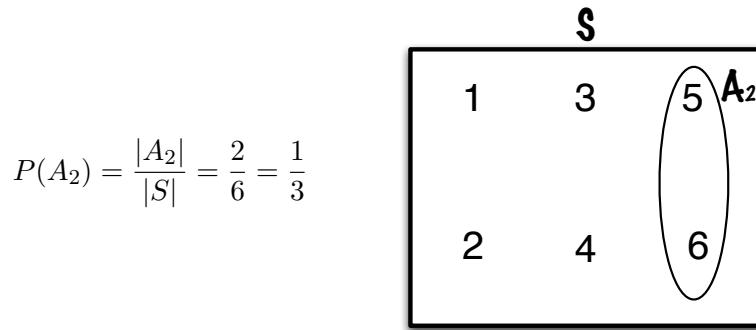
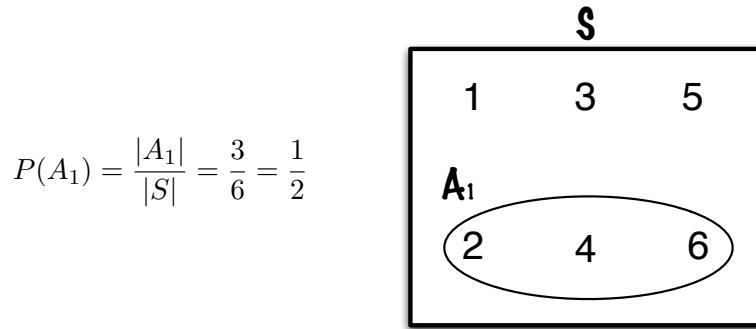
$$P(A) = \frac{|A|}{|S|},$$

where  $|A|$  denotes the number of sample points in the event  $A$ .

### 2.2.3 Example 2.3 continued

It is often useful in simple examples like this to draw a diagram (known as a “Venn diagram”) to represent the sample space, and then label specific events in the diagram by grouping together individual sample points.

$$S = \{1, 2, 3, 4, 5, 6\} \quad A_1 = \{2, 4, 6\} \quad A_2 = \{5, 6\}$$



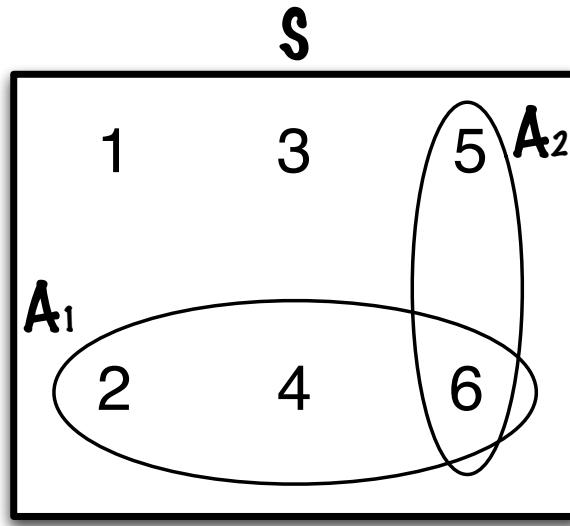
### 2.2.4 Intersection

What about  $P(\text{face is even, and larger than 4})$ ?

We can write this event in set notation as  $A_1 \cap A_2$ .

This is the **intersection** of the two events,  $A_1$  and  $A_2$   
i.e the set of elements which belong to both  $A_1$  and  $A_2$ .

$$A_1 \cap A_2 = \{6\} \quad \Rightarrow \quad P(A_1 \cap A_2) = \frac{|A_1 \cap A_2|}{|S|} = \frac{1}{6}$$



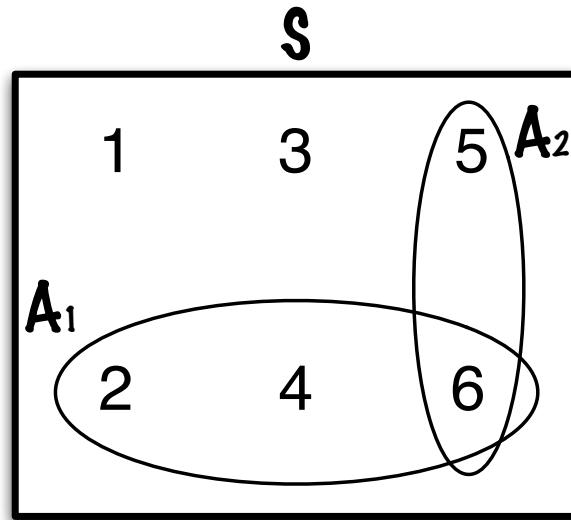
### 2.2.5 Union

What about  $P(\text{face is even, or larger than 4})$  ?

We can write this event in set notation as  $A_1 \cup A_2$ .

This is the **union** of the two events,  $A_1$  and  $A_2$   
i.e the set of elements which belong either  $A_1$  and  $A_2$  or both.

$$A_1 \cup A_2 = \{2, 4, 5, 6\} \quad \Rightarrow \quad P(A_1 \cup A_2) = \frac{|A_1 \cup A_2|}{|S|} = \frac{4}{6} = \frac{2}{3}$$



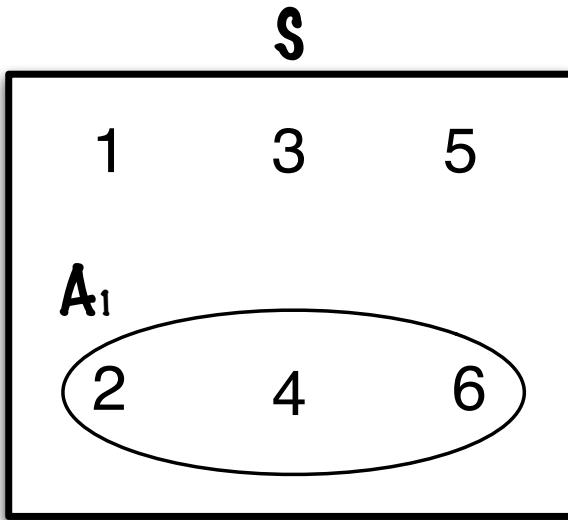
### 2.2.6 Complement

What about  $P(\text{face is not even})$  ?

We can write this event in set notation as  $A_1^c$ .

This is the **complement** of the event,  $A_1$   
i.e the set of elements which do not belong to  $A_1$ .

$$A_1^c = \{1, 3, 5\} \quad \Rightarrow \quad P(A_1^c) = \frac{|A_1^c|}{|S|} = \frac{3}{6} = \frac{1}{2}$$



## 2.3 Probability in more general settings

In many settings, either the sample space is infinite or all possible outcomes of the experiment are not equally likely. We still wish to associate probabilities with events of interest. Luckily, there are some rules/laws that allow us to calculate and manipulate such probabilities with ease.

### 2.3.1 Probability Axioms (Building Blocks)

Before we consider the probability rules we need to know about the axioms (or mathematical building blocks) upon which these rules are built. There are three axioms which we need in order to develop our laws

- (i).  $0 \leq P(A) \leq 1$  for any event  $A$ .

*This axiom says that probabilities must lie between 0 and 1*

- (ii).  $P(S) = 1$ .

*This axiom says that the probability of everything in the sample space is 1. This says that the sample space is complete and that there are no sample points or events that allow outside the sample space that can occur in our experiment.*

- (iii). If  $A_1, \dots, A_n$  are **mutually exclusive** events, then

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n).$$

A set of events are **mutually exclusive** if at most one of the events can occur in a given experiment.

*This axiom says that to calculate the probability of the union of distinct events we can simply add their individual probabilities.*

### 2.3.2 Complement Law

If  $A$  is an event, the set of all outcomes that are not in  $A$  is called the **complement** of the event  $A$ , denoted  $A^C$ .

This is (pronounced ‘A complement’).

The rule is

$$A^C = 1 - P(A) \quad (\text{Law 1})$$

#### Example 2.4: Complements

Let  $S$  (the sample space) be the set of students at Oxford. We are picking a student at random.

Let  $A$  = The event that the randomly selected student suffers from depression

We are told that 8% of students suffer from depression, so  $P(A) = 0.08$ . What is the probability that a student does not suffer from depression?

The event {student does not suffer from depression} is  $A^C$ . If  $P(A) = 0.08$  then  $P(A^C) = 1 - 0.08 = 0.92$ . ■

### 2.3.3 Addition Law (Union)

Suppose,

- A = The event that a randomly selected student from the class has brown eyes
- B = The event that a randomly selected student from the class has blue eyes

What is the probability that a student has brown eyes **OR** blue eyes?

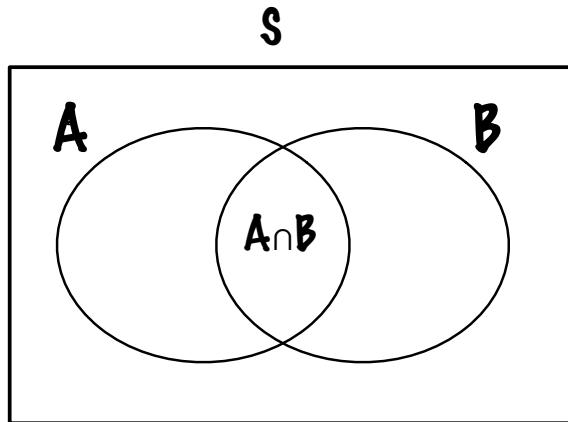
This is the **union** of the two events A and B, denoted  $A \cup B$  (pronounced ‘A or B’)

We want to calculate  $P(A \cup B)$ .

In general for two events

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{Addition Law})$$

To understand this law consider a Venn diagram of the situation (below) in which we have two events  $A$  and  $B$ . The event  $A \cup B$  is represented in such a diagram by the combined sample points enclosed by  $A$  or  $B$ . If we simply add  $P(A)$  and  $P(B)$  we will count the sample points in the intersection  $A \cap B$  twice and thus we need to subtract  $P(A \cap B)$  from  $P(A) + P(B)$  to calculate  $P(A \cup B)$ .



### Example 2.5: SNPs

Single nucleotide polymorphisms (SNPs) are nucleotide positions in a genome which exhibit variation amongst individuals in a species. In some studies in humans, SNPs are discovered in European populations. Suppose that of such SNPs, 70% also show variation in an African population, 80% show variation in an

Asian population and 60% show variation in both the African and Asian population.

Suppose one such SNP is chosen at random, what is the probability that it is variable in either the African or the Asian population?

Write  $A$  for the event that the SNP is variable in Africa, and  $B$  for the event that it is variable in Asia. We are told

$$\begin{aligned}P(A) &= 0.7 \\P(B) &= 0.8 \\P(A \cap B) &= 0.6.\end{aligned}$$

We require  $P(A \cup B)$ . From the addition rule:

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&= 0.7 + 0.8 - 0.6 \\&= 0.9.\end{aligned}$$



## Lecture 3

# Probability II

### 3.1 Independence and the Multiplication Law

If the probability that one event A occurs doesn't affect the probability that the event B also occurs, then we say that A and B are **independent**. For example, it seems clear than one coin doesn't know what happened to the other one (and if it did know, it wouldn't care), so if  $A_1$  is the event that the first coin comes up heads, and  $A_2$  the event that the second coin comes up heads, then

**Example 3.1:** One die, continued

Continuing from Example 2.3, with event  $A_1 = \{\text{face is even}\} = \{2, 4, 6\}$  and  $A_2 = \{\text{face is greater than 4}\} = \{5, 6\}$ , we see that  $A_1 \cap A_2 = \{6\}$

$$\begin{aligned} P(A_1) &= \frac{3}{6} = 0.5, \\ P(A_2) &= \frac{2}{6} = 0.33, \\ P(A_1 \cap A_2) &= \frac{1}{6} = P(A_1) \times P(A_2). \end{aligned}$$

On the other hand, if  $A_3 = \{4, 5, 6\}$ , then  $A_3$  and  $A_1$  are not independent. ■

### Example 3.2: Two dice

Suppose we roll two dice. The sample space may be represented as pairs (first roll, second roll).

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

There are 36 points in the sample space. These are all equally likely. Thus, each point has probability  $1/36$ . Consider the events

$$A = \{\text{First roll is even}\},$$

$$B = \{\text{Second roll is bigger than 4}\},$$

$$A \cap B = \{\text{First roll is even and Second roll is bigger than 4}\},$$

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
A	(2,1)	(2,2)	(2,3)	(2,4)	(2,5) (2,6)
	(3,1)	(3,2)	(3,3)	(3,4)	(3,5) (3,6)
A	(4,1)	(4,2)	(4,3)	(4,4)	(4,5) (4,6)
	(5,1)	(5,2)	(5,3)	(5,4)	(5,5) (5,6)
A	(6,1)	(6,2)	(6,3)	(6,4)	(6,5) (6,6)
				B	

Figure 3.1: Events  $A = \{\text{First roll is even}\}$  and  $B = \{\text{Second roll is bigger than 4}\}$ .

We see from Figure 3.1 that  $A$  contains 18 points and  $B$  contains 12 points, so that  $P(A) = 18/36 = 1/2$ , and  $P(B) = 12/36 = 1/3$ . Meanwhile,  $A \cap B$  contains 6 points, so  $P(A \cap B) = 6/36 = 1/6 = 1/2 \times 1/3$ . Thus  $A$  and  $B$  are independent. This should not be surprising:  $A$  depends only on the first roll, and  $B$  depends on the second. These two have no effect on each other, so the events must be independent. ■

This points up an important rule:

Events that depend on experiments that can't influence each other are always independent.

Thus, two (or more) coin flips are always independent. But this is also relevant to analysing experiments such as those of Example 2.1. If the drug has no effect on survival, then events like {patient # 612 survived} are independent of events like {patient # 612 was allocated to the control group}.

### Example 3.3: Two dice

Suppose we roll two dice. Consider the events

$$\begin{aligned} A &= \{\text{First roll is even}\}, \\ C &= \{\text{Sum of the two rolls is bigger than } 8\}, \\ A \cap C &= \{\text{First roll is even and sum is bigger than } 8\}, \end{aligned}$$

Then we see from Figure 3.2 that  $P(C) = 10/36$ , and  $P(A \cap C) = 6/36 \neq 10/36 \times 1/2$ . On the other hand, if we replace  $C$  by  $D = \{\text{Sum of the two rolls is exactly } 9\}$ , then we see from Figure 3.3 that  $P(D) = 4/36 = 1/9$ , and  $P(A \cap D) = 2/36 = 1/9 \times 1/2$ , so the events  $A$  and  $D$  are independent. We see that events may be independent, even if they are not based on separate experiments.



(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	
A	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
A	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
A	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Figure 3.2: Events  $A = \{\text{First roll is even}\}$  and  $C = \{\text{Sum is bigger than 8}\}$ .

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	
A	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
A	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
A	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Figure 3.3: Events  $A = \{\text{First roll is even}\}$  and  $D = \{\text{Sum is exactly 9}\}$ .

### Example 3.4: DNA Fingerprinting

A simplified description of how DNA fingerprinting works is this: The police find biological material connected to a crime. In the laboratory, they identify certain SNPs, finding out which version of each SNP the presumed culprit has. Then, they either search a database for someone who has the same versions of all the SNPs, or compares these SNPs to those of a suspect.

Searching a database can be potentially problematic. Imagine that the laboratory has found 12 SNPs, at which the crime-scene DNA has rare versions, each of which is found in only 10% of the population. They then search a database and find someone with all the same SNP versions. The expert then comes and testifies, to say that the probability of any single person having the same SNPs is  $(1/10)^{12} = 1/1$  trillion. There are only 60 million people in the UK, so the probability of there being another person with the same SNPs is only about 60 million/1 trillion = 0.00006 — less than 1 in ten thousand. So it can't be mistaken identity.

Except... Having particular variants at different SNPs are not independent events. For one thing, some SNPs in one population (Europeans, for example) may not be SNPs in other population (Asians, for example) where everyone may have the same variant. Thus, the 10% of the population that has each of these different rare SNP variants could in fact be the same 10%, and they may have all of these dozen variants in common because they all come from the same place, where everyone has just those variants.

And don't forget to check whether the suspect has a monozygotic twin! More than 1 person in a thousand has one, and in that case, the twins will have all the same rare SNPs, because their genomes are identical. ■

## 3.2 Conditional Probability Laws

Suppose,

- A = The event that a randomly selected student from the class has a bike  
B = The event that a randomly selected student from the class has blue eyes

and  $P(A) = 0.36$ ,  $P(B) = 0.45$  and  $P(A \cap B) = 0.13$

What is the probability that a student has a bike **GIVEN** that the student has blue eyes?

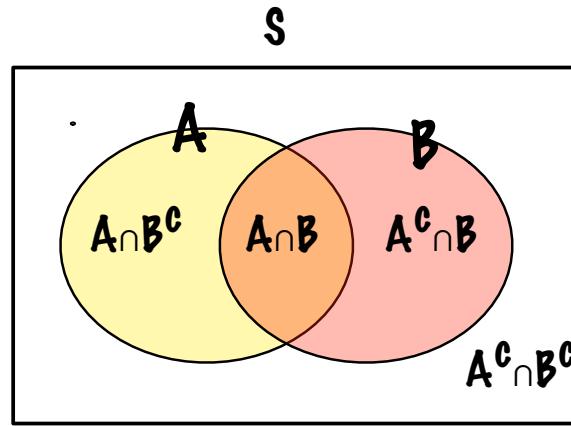
in other words

Considering just students who have blue eyes, what is the probability that a randomly selected student has a bike?

This is a **conditional** probability.

We write this probability as  $P(B|A)$  (pronounced ‘probability of B given A’)

Think of  $P(B|A)$  as ‘how much of A is taken up by B’.



Then we see that

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

(Conditional Probability Law)

### Example 3.5: SNPs again

We return to the setting of Example 2.5. What is the probability that a SNP is variable in the African population given that it is variable in the Asian population?

We have that

$$\begin{aligned} P(A) &= 0.7 \\ P(B) &= 0.8 \\ P(A \cap B) &= 0.6. \end{aligned}$$

We want

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.6}{0.8} = 0.75$$

■

We can rearrange the conditional probability law to obtain a general Multiplication Law.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(B|A)P(A) = P(A \cap B)$$

Similarly

$$P(A|B)P(B) = P(A \cap B)$$

$$\Rightarrow \boxed{P(A \cap B) = P(B|A)P(A) = P(A|B)P(B)}$$

### Example 3.6: Multiplication Law

If  $P(B) = 0.2$  and  $P(A|B) = 0.36$  what is  $P(A \cap B)$ ?

$$P(A \cap B) = 0.36 \times 0.2 = 0.072 \blacksquare$$

#### 3.2.1 Independence of Events

**Definition** Two events are  $A$  and  $B$  are said to be *independent* if  $P(A \cap B) = P(A)P(B)$ .

Note that in this case (provided  $P(B) > 0$ ), if  $A$  and  $B$  are independent

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A),$$

and similarly  $P(B|A) = P(B)$  (provided  $P(A) > 0$ ).

So for independent events, knowledge that one of the events has occurred does not change our assessment of the probability that the other event has occur.

### Example 3.7: Snails

In a population of a particular species of snail, individuals exhibit different forms. It is known that 45% have a pink background colouring, while 55% have a yellow background colouring. In addition, 30% of individuals are striped, and 20% of the population are pink and striped.

1. Is the presence or absence of striping independent of background colour?
2. Given that a snail is pink, what is the probability that it will have stripes.

Define the events:  $A$ ,  $B$ , that a snail has a pink, respectively yellow, background colouring, and  $S$  for the event that is has stripes.

Then we are told  $P(A) = 0.45$ ,  $P(B) = 0.55$ ,  $P(S) = 0.3$ , and  $P(A \cap S) = 0.2$ .

For part (1), note that

$$0.2 = P(A \cap S) \neq 0.135 = P(A)P(S),$$

so the events  $A$  and  $S$  are not independent.

For part (2),

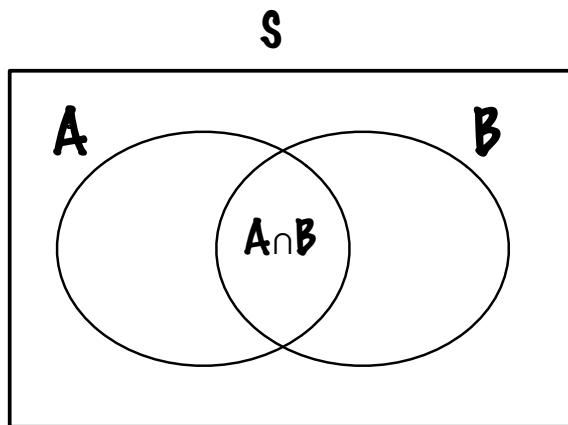
$$P(S|A) = \frac{P(S \cap A)}{P(A)} = \frac{0.2}{0.45} = 0.44.$$

Thus, knowledge that a snail has a pink background colouring increases the probability that it is striped. (That  $P(S|A) \neq P(S)$  also establishes that background colouring and the presence of stripes, are not independent.) ■

### 3.2.2 The Partition law

The partition law is a very useful rule that allows us to calculate the probability of an event by splitting it up into a number of mutually exclusive events. For example, suppose we know that  $P(A \cap B) = 0.52$  and  $P(A \cap B^c) = 0.14$  what is  $p(A)$ ?

$P(A)$  is made up of two parts (i) the part of A contained in B (ii) the part of A contained in  $B^c$ .



So we have the rule

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

$$\text{and } P(A) = P(A \cap B) + P(A \cap B^c) = 0.52 + 0.14 = 0.66$$

More generally, events are *mutually exclusive* if at most one of the events can occur in a given experiment. Suppose  $E_1, \dots, E_n$  are *mutually exclusive* events, which together form the whole sample space:  $E_1 \cup E_2 \cup \dots \cup E_n = S$ . (In other words, every possible outcome is in exactly one of the  $E$ 's. Then

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A|E_i)P(E_i)$$

### Example 3.8: Mendelian segregation

At a particular locus in humans, there are two alleles  $A$  and  $B$ , and it is known that the population frequencies of the genotypes  $AA$ ,  $AB$ , and  $BB$ , are 0.49, 0.42, and 0.09, respectively. An  $AA$  man has a child with a woman whose genotype is unknown.

What is the probability that the child will have genotype  $AB$ ?

We assume that as far as her genotype at this locus is concerned the woman is chosen randomly from the population.

Use the partition rule, where the partition corresponds to the three possible genotypes for the woman. Then

$$\begin{aligned}
 P(\text{child } AB) &= P(\text{child } AB \text{ and mother } AA) \\
 &\quad + P(\text{child } AB \text{ and mother } AB) \\
 &\quad + P(\text{child } AB \text{ and mother } BB) \\
 &= P(\text{mother } AA)P(\text{child } AB|\text{mother } AA) \\
 &\quad + P(\text{mother } AB)P(\text{child } AB|\text{mother } AB) \\
 &\quad + P(\text{mother } BB)P(\text{child } AB|\text{mother } BB) \\
 &= 0.49 \times 0 + 0.42 \times 0.5 + 0.09 \times 1 \\
 &= 0.3.
 \end{aligned}$$

■

## 3.3 Bayes' Rule

One of the most common situations in science is that we have some observations, and we need to figure out what state of the world is likely to have produced those observations. For instance, we observe that a certain number of vaccinated people contract polio, and a certain number of unvaccinated people contract polio, and we need to figure out how effective the vaccine is. The problem is, our theoretical knowledge goes in the wrong direction: If the vaccine is so effective, how many people will contract polio. Bayes' Rule allows us to turn the inference around.

**Example 3.9: Medical testing**

In a medical setting we might want to calculate the probability that a person has a disease D given they have a specific symptom S, i.e. we want to calculate  $P(D|S)$ . This is a hard probability to assign as we would need to take a random sample of people from the population with the symptom.

A probability that is much easier to calculate is  $P(S|D)$ , i.e. the probability that a person with the disease has the symptom. This probability can be assigned much more easily as medical records for people with serious diseases are kept.

The power of Bayes Rule is its ability to take  $P(S|D)$  and calculate  $P(D|S)$ .

*We have already seen a version of Bayes' Rule*

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

*Using the Multiplication Law we can rewrite this as*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (\text{Bayes Rule})$$

Suppose  $P(S|D) = 0.12$ ,  $P(D) = 0.01$  and  $P(S) = 0.03$ . Then

$$P(D|S) = \frac{0.12 \times 0.01}{0.03} = 0.04$$

■

**Example 3.10: Genetic testing**

A gene has two possible types  $A_1$  and  $A_2$ . 75% of the population have  $A_1$ .  $B$  is a disease that has 3 forms  $B_1$  (mild),  $B_2$  (severe)

and  $B_3$  (lethal).  $A_1$  is a protective gene, with the probabilities of having the three forms given  $A_1$  as 0.9, 0.1 and 0 respectively. People with  $A_2$  are unprotected and have the three forms with probabilities 0, 0.5 and 0.5 respectively.

What is the probability that a person has gene  $A_1$  given they have the severe disease?

The first thing to do with such a question is ‘decode’ the information, i.e. write it down in a compact form we can work with.

$$P(A_1) = 0.75 \quad P(A_2) = 0.25$$

$$P(B_1|A_1) = 0.9 \quad P(B_2|A_1) = 0.1 \quad P(B_3|A_1) = 0$$

$$P(B_1|A_2) = 0 \quad P(B_2|A_2) = 0.5 \quad P(B_3|A_2) = 0.5$$

We want  $P(A_1|B_2)$ ?

From Bayes Rule we know that

$$P(A_1|B_2) = \frac{P(B_2|A_1)P(A_1)}{P(B_2)}$$

We know  $P(B_2|A_1)$  and  $P(A_1)$  but what is  $P(B_2)$ ?

We can use the Partition Law since  $A_1$  and  $A_2$  are mutually exclusive.

$$\begin{aligned} P(B_2) &= P(B_2 \cap A_1) + P(B_2 \cap A_2) && \text{(Partition Law)} \\ &= P(B_2|A_1)P(A_1) + P(B_2|A_2)P(A_2) && \text{(Multiplication Law)} \\ &= 0.1 \times 0.75 + 0.5 \times 0.25 \\ &= 0.2 \end{aligned}$$

We can use Bayes Rule to calculate  $P(A_1|B_2)$ .

$$P(A_1|B_2) = \frac{0.1 \times 0.75}{0.2} = 0.375$$



### 3.4 Probability Laws

$$P(A^c) = 1 - P(A) \quad (\text{Complement Law})$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{Addition Law})$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (\text{Conditional Probability Law})$$

$$P(A \cap B) = P(B|A)P(A) = P(A|B)P(B) \quad (\text{Multiplication Law})$$

If  $E_1, \dots, E_n$  are a set of *mutually exclusive* events then

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A|E_i)P(E_i) \quad (\text{Partition Law})$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (\text{Bayes Rule})$$

### 3.5 Permutations and Combinations (Probabilities of patterns)

In some situations we observe a specific pattern from a large number of possible patterns. To calculate the probability of the pattern we need to count the total number of patterns. This is why we need to learn about permutations and combinations.

#### 3.5.1 Permutations of $n$ objects

Consider 2 objects    A              B

Q. How many ways can they be arranged? i.e. how many **permutations** are there?

A. 2 ways      AB      BA

Consider 3 objects    A            B            C

Q. How many ways can they be arranged (permuted)?

A. 6 ways      ABC    ACB    BCA    BAC    CAB    CBA

Consider 4 objects    A            B            C            D

Q. How many ways can they be arranged (permuted)?

A. 24 ways

ABCD	ABDC	ACBD	ACDB	ADBC	ADCB
BACD	BADC	BCAD	BCDA	BDAC	BDCA
CBAD	CBDA	CABD	CADB	CDBA	CDAB
DBCA	DBAC	DCBA	DCAB	DABC	DACB

There is a pattern emerging here.

No. of objects	2	3	4	5	6	...
No. of permutations	2	6	24	120	720	...

Can we find a formula for the number of permutations of  $n$  objects?

A good way to think about permutations is to think of putting objects into boxes.

Suppose we have 5 objects. How many different ways can we place them into 5 boxes?



There are 5 choices of object for the first box.



There are now only 4 objects to choose from for the second box.



There are 3 choices for the 3rd box, 2 for the 4th and 1 for the 5th box.

5	4	3	2	1
---	---	---	---	---

Thus, the number of permutations of 5 objects is  $5 \times 4 \times 3 \times 2 \times 1$ .

In general, the number of permutations of  $n$  objects is

$$n(n-1)(n-2)\dots(3)(2)(1)$$

We write this as  $n!$  (pronounced ‘n factorial’). There should be a button on your calculator that calculates factorials.

### 3.5.2 Permutations of $r$ objects from $n$

Now suppose we have 4 objects and only 2 boxes. How many permutations of 2 objects when we have 4 to choose from?

There are 4 choices for the first box and 3 choices for the second box

4	3
---	---

So there are 12 permutations of 2 objects from 4. We write this as

$${}^4P_2 = 12$$

We say there are  ${}^nP_r$  permutations of  $r$  objects chosen from  $n$ .

The formula for  ${}^nP_r$  is given by

$${}^nP_r = \frac{n!}{(n-r)!}$$

To see why this works consider the example above  ${}^4P_2$ .

Using the formula we get

$${}^4P_2 = \frac{4!}{2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 4 \times 3$$

### 3.5.3 Combinations of $r$ objects from $n$

Now consider the number of ways of choosing 2 objects from 4 when the order doesn't matter. We just want to count the number of possible **combinations**.

We know that there are 12 permutations when choosing 2 objects from 4. These are

$$\begin{array}{ccccccc} AB & AC & AD & BC & BD & CD \\ BA & CA & DA & CB & DB & DC \end{array}$$

Notice how the permutations are grouped in 2's which are the same combination of letters. Thus there are  $12/2 = 6$  possible combinations.

$$AB \quad AC \quad AD \quad BC \quad BD \quad CD$$

We write this as

$${}^4C_2 = 6$$

We say there are  ${}^nC_r$  combinations of  $r$  objects chosen from  $n$ .

The formula for  ${}^nC_r$  is given by

$${}^nC_r = \frac{n!}{(n-r)!r!}$$

Another way of writing this formula that makes it clearer is

$${}^nC_r = \frac{{}^nP_r}{r!}$$

Effectively this says we count the number of permutations of  $r$  objects from  $n$  and then divide by  $r!$  because the  ${}^nP_r$  permutations will occur in groups of  $r!$  that are the same combination.

### 3.6 Worked Examples

- (i). Four letters are chosen at random from the word RANDOMLY. Find the probability that all four letters chosen are consonants.

8 letters, 6 consonants, 2 vowels

$$P(\text{all four are consonants}) = \frac{\# \text{ of ways of choosing 4 consonants}}{\# \text{ of ways of choosing 4 letters}}$$

$$\# \text{ of ways of choosing 4 consonants} = {}^6C_4 = \frac{6!}{4!2!} = 15$$

$$\# \text{ of ways of choosing 4 letters} = {}^8C_4 = \frac{8!}{4!4!} = 70$$

$$\Rightarrow P(\text{all four are consonants}) = \frac{15}{70} = \frac{3}{14}$$

- (ii). A bag contains 8 white counters and 3 black counters. Two counters are drawn, one after the other. Find the probability of drawing one white and one black counter, in any order

- (a) if the first counter is replaced
- (b) if the first counter is not replaced

What is the probability that the second counter is black (assume that the first counter is replaced after it is taken)?

*A useful way of tackling many probability problems is to draw a ‘probability tree’. The branches of the tree represent different possible events. Each branch is labelled with the probability of choosing it given what has occurred before. The probability of a given route through the tree can then be calculated by multiplying all the probabilities along that route (using the Multiplication Rule)*

- (a) *With replacement*

Let

$W_1$  be the event ‘a white counter is drawn first’,

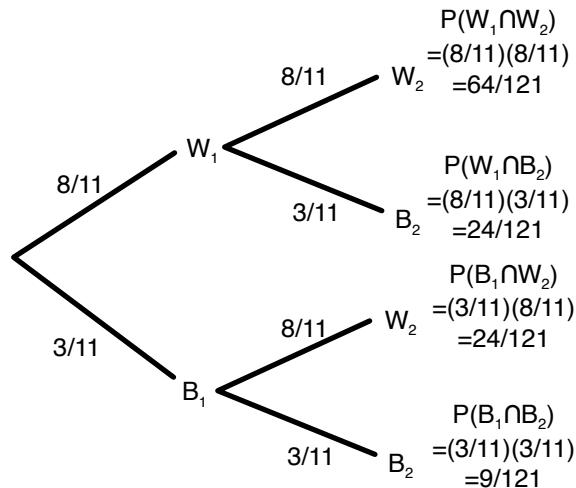
$W_2$  be the event ‘a white counter is drawn second’,

$B_1$  be the event ‘a black counter is drawn first’,

$B_2$  be the event ‘a black counter is drawn second’,

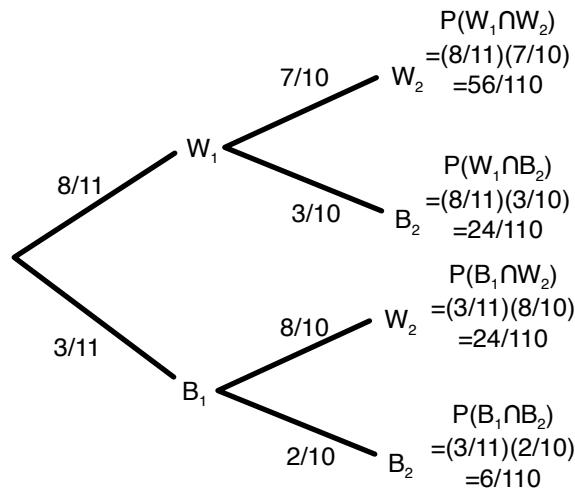
$$\begin{aligned}
 P(\text{one white and one black counter}) &= P(W_1 \cap B_2) + P(B_1 \cap W_2) \\
 &= \frac{24}{121} + \frac{24}{121} \\
 &= \frac{48}{121}
 \end{aligned}$$

(b) *Without replacement*



$$\begin{aligned}
 P(\text{one white and one black counter}) &= P(W_1 \cap B_2) + P(B_1 \cap W_2) \\
 &= \frac{24}{110} + \frac{24}{110} \\
 &= \frac{48}{110}
 \end{aligned}$$

$$\begin{aligned}
 P(\text{second counter is black}) &= P(W_1 \cap B_2) + P(B_1 \cap B_2) \\
 &= \frac{24}{121} + \frac{9}{121} \\
 &= \frac{33}{121}
 \end{aligned}$$



(iii). **From 2001 TT Prelim Q1** Two drugs that relieve pain are available to treat patients. Drug A has been found to be effective in three-quarters of all patients; when it is effective, the patients have relief from pain one hour after taking this drug. Drug B acts quicker but only works with one half of all patients: those who benefit from this drug have relief of pain after 30 mins. The physician cannot decide which patients should be prescribed which drug so he prescribes randomly. Assuming that there is no variation between patients in the times taken to act for either drug, calculate the probability that:

- (a) a patient is prescribed drug B and is relieved of pain;
- (b) a patient is relieved of pain after one hour;
- (c) a patient who was relieved of pain after one hour took drug A;
- (d) two patients receiving different drugs are both relieved of pain after one hour.
- (e) out of six patients treated with the same drug, three are relieved of pain after one hour and three are not.

Let

$R_{30}$  = The event that a patient is relieved of pain within 30 mins

$R_{60}$  = The event that a patient is relieved of pain within 60 mins

$A$  = Event that a patient takes drug A

$B$  = Event that a patient takes drug B

$$P(R60|A) = 0.75 \quad P(R30|B) = 0.5 \quad P(A) = P(B) = 0.5$$

$$(a) P(R30|B)P(B) = 0.25$$

$$(b) P(R60) = P(R60|A)P(A) + P(R60|B)P(B) \text{ since R30} \Rightarrow \text{R60}$$

$$P(R60) = 0.75 \times 0.5 + 0.5 \times 0.5 = 0.625$$

$$(c) P(A|R60) = \frac{P(A \cap R60)}{P(R60)}$$

$$P(A \cap R60) = P(R60|A)P(A) = 0.75 \times 0.5 = 0.375$$

$$\Rightarrow P(A|R60) = \frac{0.375}{0.625} = 0.6$$

$$(d) P(R60|A)P(R60|B) = 0.75 \times 0.5 = 0.375$$

Assuming the events are independent.

$$(e) n = 6 \quad X = \text{no. of patients relieved of pain after 1hr}$$

$$\text{For } A, p = P(R60|A) = 0.75$$

$$P(X = 3|A) = {}^6C_3(0.75)^3(0.25)^3 = 0.1312$$

$$\text{For } B, p = P(R60|B) = 0.5$$

$$P(X = 3|B) = {}^6C_3(0.5)^3(0.5)^3 = 0.3125$$

$$\Rightarrow P(X = 3) = P(X = 3|A)P(A) + (X = 3|B)P(B)$$

$$= 0.1312 \times 0.5 + 0.3125 \times 0.5 = 0.2222$$

4. In the National Lottery you need to choose 6 balls from 49.

What is the probability that I choose all 6 balls correctly?

There are 2 ways of answering this question

- (i) using permutations and combinations
- (ii) using a tree diagram

**Method 1 - using permutations and combinations**

$$\begin{aligned}
 P(6 \text{ correct}) &= \frac{\text{No. of ways of choosing the 6 correct balls}}{\text{No. of ways of choosing 6 balls}} \\
 &= \frac{^6P_6}{49P_6} \\
 &= \frac{\frac{6!}{0!}}{\frac{49!}{43!}} \\
 &= \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{49 \times 48 \times 47 \times 46 \times 45 \times 44} \\
 &= 0.0000000715112 \quad (\text{1 in 14 million})
 \end{aligned}$$

**Method 2 - using a tree diagram**

Consider the first ball I choose, the probability it is correct is

$$\frac{6}{49}$$

The second ball I choose is correct with probability

$$\frac{5}{48}$$

The third ball I choose is correct with probability

$$\frac{4}{47}$$

and so on.

Thus the probability that I get all 6 balls correct is

$$\frac{6}{49} \frac{5}{48} \frac{4}{47} \frac{3}{46} \frac{2}{45} \frac{1}{44} = 0.0000000715112 \quad (\text{1 in 14 million})$$



## Lecture 4

# The Binomial Distribution

### 4.1 Introduction

In Lecture 2 we saw that we need to study probability so that we can calculate the ‘chance’ that our sample leads us to the wrong conclusion about the population. To do this in practice we need to ‘model’ the process of taking the sample from the population. By ‘model’ we mean describe the process of taking the sample in terms of the probability of obtaining each possible sample. Since there are many different types of data and many different ways we might collect a sample of data we need lots of different probability models. The Binomial distribution is one such model that turns out to be very useful in many experimental settings.

### 4.2 An example of the Binomial distribution

Suppose we have a box with a very large number<sup>1</sup> of balls in it:  $\frac{2}{3}$  of the balls are black and the rest are red. We draw 5 balls from the box. How many black balls do we get? We can write

$$X = \text{No. of black balls in 5 draws.}$$

$X$  can take on any of the values 0, 1, 2, 3, 4 and 5.

$X$  is a **discrete random variable**

---

<sup>1</sup>We say “a very large number” when we want to ignore the change in probability that comes from drawing without replacement. Alternatively, we could have a small number of balls — 2 black and 1 red, for instance — but replace the ball (and mix well!) after each draw.

Some values of X will be more likely to occur than others. Each value of X will have a probability of occurring. What are these probabilities? Consider the probability of obtaining just one black ball, i.e. X = 1.

One possible way of obtaining one black ball is if we observe the pattern BR<sub>4</sub>R. The probability of obtaining this pattern is

$$P(BR\bar{R}R) = \frac{2}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3}$$

There are 32 possible patterns of black and red balls we might observe. 5 of the patterns contain just one black ball

BBBBB	RBBBB	BRBBB	BBRBB	BBBRB	BBBBR	RRBBB	RBRBB
RBBRB	RBBBR	BRRBB	BRBRB	BRBBR	BBRRB	BBRBR	BBBRR
RRRBB	RRRBR	RRBBR	RBRBB	RBRBR	RBBRR	BRRRB	BRRBR
BRBRR	BBRRR	<span style="border: 1px solid black; padding: 2px;">BRRRR</span>	<span style="border: 1px solid black; padding: 2px;">RBRRR</span>	<span style="border: 1px solid black; padding: 2px;">RRBRR</span>	<span style="border: 1px solid black; padding: 2px;">RRRBR</span>	<span style="border: 1px solid black; padding: 2px;">RRRRB</span>	RRRRR

The other 5 possible combinations all have the same probability so the probability of obtaining one head in 5 coin tosses is

$$P(X = 1) = 5 \times \left( \frac{2}{3} \times \left( \frac{1}{3} \right)^4 \right) = 0.0412 \text{ (to 4 decimal places)}$$

What about P(X = 2)? This probability can be written as

$$\begin{aligned} P(X = 2) &= \text{No. of patterns} \times \text{Probability of pattern} \\ &= {}^5C_2 \times \left( \frac{2}{3} \right)^2 \times \left( \frac{1}{3} \right)^3 \\ &= 10 \times \frac{4}{243} \\ &= 0.165 \end{aligned}$$

It's now just a small step to write down a formula for this situation specific situation in which we toss a coin 5 times

$$P(X = x) = {}^5C_x \times \left( \frac{2}{3} \right)^x \times \left( \frac{1}{3} \right)^{(5-x)}$$

We can use this formula to tabulate the probabilities of each possible value of X.

These probabilities are plotted in Figure 4.1 against the values of X. This shows the **distribution** of probabilities across the possible values of X. This

$$\begin{aligned}
 P(X = 0) &= {}^5C_0 \times \left(\frac{2}{3}\right)^0 \times \left(\frac{1}{3}\right)^5 = 0.0041 \\
 P(X = 1) &= {}^5C_1 \times \left(\frac{2}{3}\right)^1 \times \left(\frac{1}{3}\right)^4 = 0.0412 \\
 P(X = 2) &= {}^5C_2 \times \left(\frac{2}{3}\right)^2 \times \left(\frac{1}{3}\right)^3 = 0.1646 \\
 P(X = 3) &= {}^5C_3 \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^2 = 0.3292 \\
 P(X = 4) &= {}^5C_4 \times \left(\frac{2}{3}\right)^4 \times \left(\frac{1}{3}\right)^1 = 0.3292 \\
 P(X = 5) &= {}^5C_5 \times \left(\frac{2}{3}\right)^5 \times \left(\frac{1}{3}\right)^0 = 0.1317
 \end{aligned}$$

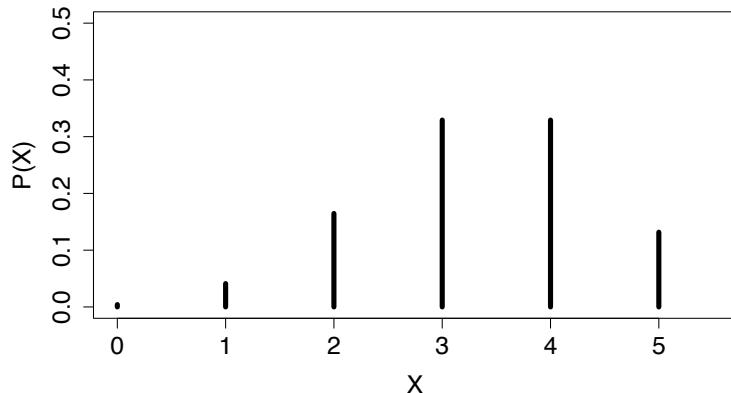


Figure 4.1: A plot of the Binomial(5, 2/3) probabilities.

situation is a specific example of a Binomial distribution.

**Note** It is important to make a distinction between the probability distribution shown in Figure 4.1 and the histograms of specific datasets seen in Lecture 1. A probability distribution represents the distribution of values we ‘expect’ to see in a sample. A histogram is used to represent the distribution of values that actually occur in a given sample.

### 4.3 The Binomial distribution

#### The key components of a Binomial distribution

In general a Binomial distribution arises when we have the following 4 conditions

- $n$  identical trials, e.g. 5 coin tosses
- 2 possible outcomes for each trial “success” and “failure”, e.g. Heads or Tails
- Trials are independent, e.g. each coin toss doesn’t affect the others
- $P(\text{“success"}) = p$  is the same for each trial, e.g.  $P(\text{Black}) = 2/3$  is the same for each trial

#### Binomial distribution probabilities

If we have the above 4 conditions then if we let

$$X = \text{No. of “successes”}$$

then the probability of observing  $x$  successes out of  $n$  trials is given by

$$P(X = x) = {}^n C_x p^x (1 - p)^{(n-x)} \quad x = 0, 1, \dots, n$$

If the probabilities of  $X$  are distributed in this way, we write

$$X \sim \text{Bin}(n, p)$$

$n$  and  $p$  are called the **parameters** of the distribution. We say  $X$  follows a binomial distribution with parameters  $n$  and  $p$ .

#### Examples

With this general formula we can calculate many different probabilities.

- (i). Suppose  $X \sim \text{Bin}(10, 0.4)$ , what is  $P(X = 7)$ ?

$$\begin{aligned} P(X = 7) &= {}^{10} C_7 (0.4)^7 (1 - 0.4)^{(10-7)} \\ &= (120)(0.4)^7 (0.6)^3 \\ &= 0.0425 \end{aligned}$$

(ii). Suppose  $Y \sim \text{Bin}(8, 0.15)$ , what is  $P(Y < 3)$ ?

$$\begin{aligned} P(Y < 3) &= P(Y = 0) + P(Y = 1) + P(Y = 2) \\ &= {}^8C_0(0.15)^0(0.85)^8 + {}^8C_1(0.15)^1(0.85)^7 + {}^8C_2(0.15)^2(0.85)^6 \\ &= 0.2725 + 0.3847 + 0.2376 \\ &= 0.8948 \end{aligned}$$

(iii). Suppose  $W \sim \text{Bin}(50, 0.12)$ , what is  $P(W > 2)$ ?

$$\begin{aligned} P(W > 2) &= P(W = 3) + P(W = 4) + \dots + P(W = 50) \\ &= 1 - P(W \leq 2) \\ &= 1 - (P(W = 0) + P(W = 1) + P(W = 2)) \\ &= 1 - ({}^{50}C_0(0.12)^0(0.88)^{50} + {}^{50}C_1(0.12)^1(0.88)^{49} + {}^{50}C_2(0.12)^2(0.88)^{48}) \\ &= 1 - (0.00168 + 0.01142 + 0.03817) \\ &= 0.94874 \end{aligned}$$

#### 4.4 The mean and variance of the Binomial distribution

Different values of  $n$  and  $p$  lead to different distributions with different shapes (see Figure 4.2). In Lecture 1 we saw that the mean and standard deviation can be used to summarize the shape of a dataset. In the case of a probability distribution we have no data as such so we must use the probabilities to calculate the *expected* mean and standard deviation. In other words, the mean and standard deviation of a random variable is the mean and standard deviation that a collection of data would have if the numbers appeared in exactly the proportions given by the distribution. The mean of a distribution is also called the **expectation** or **expected value** of the distribution.

Consider the example of the Binomial distribution we saw above

x	0	1	2	3	4	5
P(X = x)	0.004	0.041	0.165	0.329	0.329	0.132

The expected mean value of the distribution, denoted  $\mu$  can be calculated as

$$\begin{aligned} \mu &= 0 \times (0.004) + 1 \times (0.041) + 2 \times (0.165) + 3 \times (0.329) + 4 \times (0.329) + 5 \times (0.132) \\ &= 3.333 \end{aligned}$$

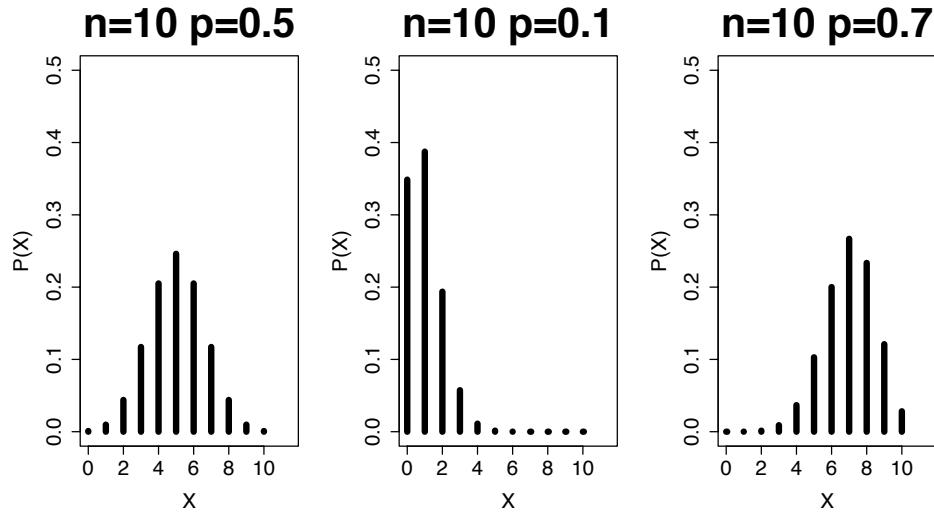


Figure 4.2: 3 different Binomial distributions.

In general, there is a formula for the mean of a Binomial distribution. There is also a formula for the standard deviation,  $\sigma$ .

If  $X \sim \text{Bin}(n, p)$  then

$$\begin{aligned}\mu &= np \\ \sigma &= \sqrt{npq} \quad \text{where } q = 1 - p\end{aligned}$$

In the example above,  $X \sim \text{Bin}(5, 2/3)$  and so the mean and standard deviation are given by

$$\mu = np = 5 \times (2/3) = 3.333$$

and

$$\sigma = \sqrt{npq} = 5 \times (2/3) \times (1/3) = 1.111$$

### Shapes of Binomial distributions

The skewness of a Binomial distribution will also depend upon the values of  $n$  and  $p$  (see Figure 4.2). In general,

- if  $p < 0.5$  the distribution will exhibit POSITIVE SKEW

- if  $p = 0.5$  the distribution will be SYMMETRIC
- if  $p > 0.5$  the distribution will exhibit NEGATIVE SKEW

However, for a given value of  $p$ , the skewness goes down as  $n$  increases. All binomial distributions eventually become approximately symmetric for large  $n$ . This will be discussed further in Lecture 6.

## 4.5 Testing a hypothesis using the Binomial distribution

Consider the following simple situation: You have a six-sided die, and you have the impression that it's somehow been weighted so that the number 1 comes up more frequently than it should. How would you decide whether this impression is correct? You could do a careful experiment, where you roll the die 60 times, and count how often the 1 comes up.

Suppose you do the experiment, and the 1 comes up 30 times (and other numbers come up 30 times all together). You might expect the 1 to come up one time in six, so 10 times, so 30 times seems high. But is it too high? There are two possible hypotheses:

- (i). The die is biased.
- (ii). Just by chance we got more 1's than expected.

How do we decide between these hypotheses? Of course, we can never prove that any sequence of throws *couldn't* have come from a fair die. But we can find that the results we got are extremely unlikely to have arisen from a fair die, so that we should seriously consider whether the alternative might be true.

Since the probability of a 1 on each throw is  $1/6$ , so we apply the formula for the binomial distribution with  $n = 60$  and  $p = 1/6$ . Then we have

Now we summarise the general approach:

- posit a **hypothesis**
- design and carry out an **experiment** to collect a **sample** of data
- **test** to see if the sample is consistent with the hypothesis

**Hypothesis** The die is **fair**. All 6 outcomes have the same probability.

**Experiment** We roll the die.

**Sample** We obtain 60 outcomes of a die roll.

**Testing the hypothesis** Assuming our hypothesis is true what is the probability that we would have observed such a sample or a sample more extreme, i.e. is our sample quite unlikely to have occurred under the assumptions of our hypothesis?

Assuming our hypothesis is true the experiment we carried out satisfies the conditions of the Binomial distribution

- $n$  identical trials, i.e. 60 die rolls.
- 2 possible outcomes for each trial: “1” and “not 1”.
- Trials are independent.
- $P(\text{“success”}) = p$  is the same for each trial, i.e.  $P(1 \text{ comes up}) = 1/6$  is the same for each trial

We define  $X = \text{No. of 1's that come up}$

We observed  $X = 30$ . Which samples are more extreme than this?

Under our hypothesis we would expect  $X = 10$

$X \geq 30$  are the samples as or more extreme than  $X = 30$ .

We can calculate each of these probabilities using the Binomial probability formula

$$P(\#\text{ 1's is exactly } 30) = {}^{60}C_{30} - \left(\frac{1}{6}\right)^{18} \left(\frac{5}{6}\right)^{60-30} = 2.25 \times 10^{-9}.$$

$$P(\#\text{ 1's is at least } 30) = \sum_{x=30}^{60} {}^{60}C_x - \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{60-x} = 2.78 \times 10^{-9}.$$

Which is the appropriate probability? The “strange event” from the perspective of the fair die was not that 1 came up exactly 30 times, but that it came up *so many* times. So the relevant number is the second one, which is a little bigger. Still, the probability is less than 3 in a billion. In other

words, if you were to perform one of these experiments once a second, continuously, you might expect to see a result this extreme once in 10 years. So you either have to believe that you just happened to get that one in 10 years outcome the one time you tried it, or you have to believe that there really is something biased about the die. In the language of hypothesis testing we say we would ‘reject the hypothesis’.

### Example 4.1: Analysing the Anturane Reinfarction Trial

From Table 2.1, we know that there were 163 patients who died, out of a total of 1629. Now, suppose the study works as follows: Patients come in the door, we flip a coin, and allocate them to the treatment group if heads comes up, or to the control group if tails comes up. (This isn’t exactly how it was done, but close enough. Next term, we’ll talk about other ways of getting the same results.)

We had a total of 813 heads out of 1629, which is pretty close to half, which seems reasonable. On the other hand, if we look at the 163 coin flips for the patients who died, we only had 74 heads, which seems pretty far from half (which would be 81.5). It seems there are two plausible hypotheses:

- (i). Anturane works. In that case, it’s perfectly reasonable to expect that fewer patients who died got the anturane treatment.
- (ii). Purely by chance we had fewer heads than expected.

One way of thinking about this formally is to use Bayes’ rule:

$$\begin{aligned} P(\text{coin heads} \mid \text{patient died}) &= P(\text{coin heads}) \frac{P(\text{patient died} \mid \text{coin heads})}{P(\text{patient died})} \\ &= \frac{1}{2} \times \frac{P(\text{patient died} \mid \text{anturane treatment})}{P(\text{patient died})}. \end{aligned}$$

If the conditional probability of dying is lowered by anturane, then retrospectively the coin flips for the patients who died have less than probability 1/2 of coming up heads; but if anturane has no effect, then these are fair coin flips, and should have come out 50% heads.

So how do we figure out which possibility is true? Of course, we can never conclusively rule out the possibility of hypothesis 2. Any number of heads is **possible**. But we can say that some numbers are extremely unlikely, indicating that it would be advisable to accept hypothesis 1 — anturane works — rather than believe that we had gotten such an exceptional result from the coin flips.

So... 74 heads in 163 flips is not the most likely outcome. But how unlikely is it? Let's consider three different probabilities:

$$\begin{aligned} P(\# \text{ heads is exactly } 74) \\ P(\# \text{ heads is at most } 74) \\ P(\# \text{ heads is at most } 74 \text{ or at least } 89). \end{aligned}$$

Which is the probability we want? It's pretty clearly not the first one. After all, any particular number of heads is pretty unlikely (and getting exactly 50% heads is impossible, since the number of tosses was odd). And if we had gotten 73 heads, that would have been considered even better evidence for hypothesis 1.

Choosing between the other two probabilities isn't so clear, though. After all, if we want to answer the question "How likely would it be to get such a strange outcome purely by chance?" we probably should consider all outcomes that would have seemed equally "strange", and 89 is as far away from the "expected" number of heads as 74. There isn't a definitive answer to choosing between these "one-tailed" and "two-tailed" tests, but we will have more to say about this later in the course.

Now we compute the probabilities:

$$P(\# \text{ heads is exactly } 74) = {}^{163}C_{74} \left(\frac{1}{2}\right)^{74} \left(\frac{1}{2}\right)^{163-74} = 0.0314,$$

$$\begin{aligned} P(\# \text{ heads is at most } 74) &= \sum_{i=0}^{74} {}^{163}C_i \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{163-i} \\ &= {}^{163}C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^{163-0} + {}^{163}C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^{163-1} \\ &\quad + \cdots + {}^{163}C_{74} \left(\frac{1}{2}\right)^{74} \left(\frac{1}{2}\right)^{163-74} \\ &= 0.136, \end{aligned}$$

$$\begin{aligned}
 P(\# \text{ heads at most } 74 \text{ or at least } 89) &= \sum_{i=0}^{74} {}^{163}C_i \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{163-i} \\
 &\quad + \sum_{i=89}^{163} {}^{163}C_i \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{163-i} \\
 &= 0.272.
 \end{aligned}$$

Note that the “two-tailed” probability is exactly twice the “one-tailed”. We show these probabilities on the probability histogram of Figure 4.3.

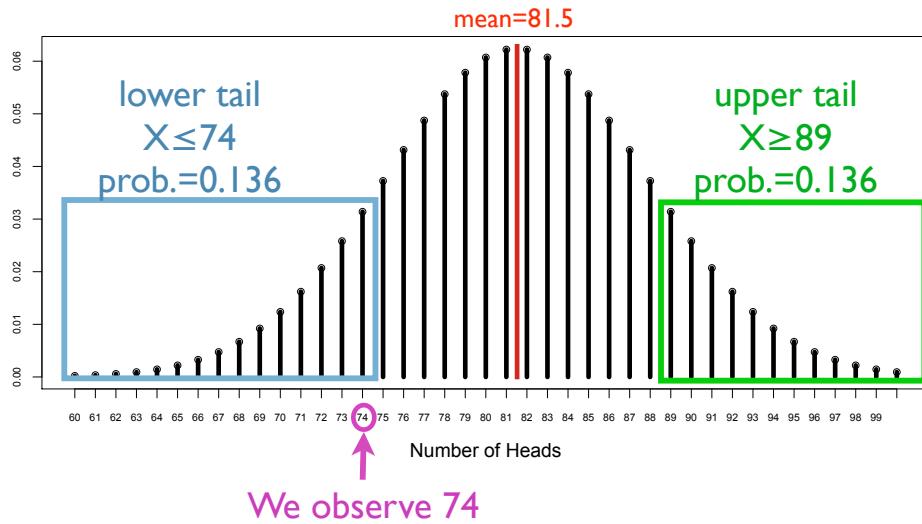


Figure 4.3: The tail probabilities for testing the hypothesis that anturane has no effect.

What is the conclusion? Even 0.136 doesn't seem like such a small probability. This says that we would expect, even if anturane does nothing at all, that we would see such a strong apparent effect about one time out of seven. Our conclusion is that this experiment does not provide significant evidence that anturane is effective. In the language of hypotheses testing we say that we *do not reject* the hypothesis that anturane is ineffective.

■



## Lecture 5

# The Poisson Distribution

### 5.1 Introduction

#### Example 5.1: Drownings in Malta

The book [Mou98] cites data from the St. Luke's Hospital Gazette, on the monthly number of drownings on Malta, over a period of nearly 30 years (355 consecutive months). Most months there were no drownings. Some months there was one person who drowned. One month had four people drown. The data are given as counts of the number of months in which a given number of drownings occurred, and we repeat them here as Table 5.1.

Looking at the data in Table 5.1, we might suppose that one of the following hypotheses is true:

- Some months are particularly dangerous;
- Or, on the contrary, when one person has drowned, the surrounding publicity makes others more cautious for a while, preventing drownings?
- Or, drownings are simply independent events?

How can we use the data to decide which of these hypotheses is true? We might reasonably suppose that the first hypothesis would predict that there would be more months with high numbers of drownings than the independence hypothesis; the second

Table 5.1: Monthly counts of drownings in Malta.

No. of drowning deaths per month	Frequency (No. months observed)
0	224
1	102
2	23
3	5
4	1
5+	0

hypothesis would predict fewer months with high numbers of drownings. The problem is, we don't know how many we should expect, if independence is correct.

What we need is a **model**: A sensible probability distribution, giving the probability of a month having a certain number of drownings, under the independence assumption. The standard model for this sort of situation is called the **Poisson distribution**. ■

The Poisson distribution is used in situations when we observe the counts of events within a set unit of time, area, volume, length etc. For example,

- The number of cases of a disease in different towns;
- The number of mutations in given regions of a chromosome;
- The number of dolphin pod sightings along a flight path through a region;
- The number of particles emitted by a radioactive source in a given time;
- The number of births per hour during a given day.

In such situations we are often interested in whether the events occur randomly in time or space. Consider the Babyboom dataset (Table 1.2), that we saw in Lecture 1. The birth times of the babies throughout the day are shown in Figure 5.1(a). If we divide up the day into 24 hour intervals and

count the number of births in each hour we can plot the counts as a histogram in Figure 5.1(b). How does this compare to the histogram of counts for a process that isn't random? Suppose the 44 birth times were distributed in time as shown in Figure 5.1(c). The histogram of these birth times per hour is shown in Figure 5.1(d). We see that the non-random clustering of events in time causes there to be more hours with zero births and more hours with large numbers of births than the real birth times histogram.

This example illustrates that the distribution of counts is useful in uncovering whether the events might occur randomly or non-randomly in time (or space). Simply looking at the histogram isn't sufficient if we want to ask the question whether the events occur randomly or not. To answer this question we need a probability model for the distribution of counts of random events that dictates the type of distributions we should expect to see.

## 5.2 The Poisson Distribution

The Poisson distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time (or space).

If we let  $X$  = The number of events in a given interval,

Then, if the mean number of events per interval is  $\lambda$

The probability of observing  $x$  events in a given interval is given by

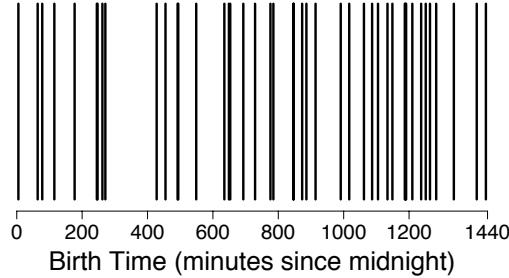
$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad x = 0, 1, 2, 3, 4, \dots$$

**Note**  $e$  is a mathematical constant.  $e \approx 2.718282$ . There should be a button on your calculator  $[e^x]$  that calculates powers of  $e$ .

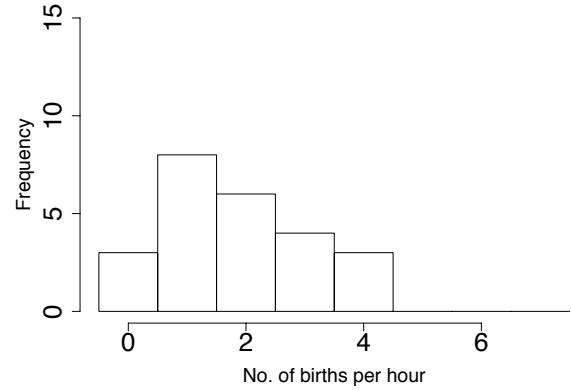
If the probabilities of  $X$  are distributed in this way, we write

$$X \sim Po(\lambda)$$

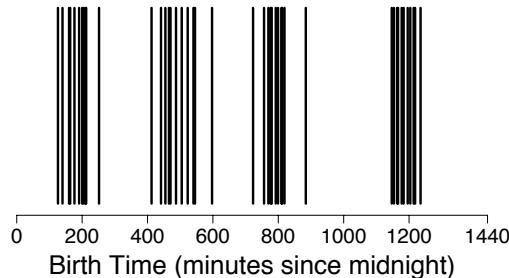
$\lambda$  is the **parameter** of the distribution. We say  $X$  follows a Poisson distribution with parameter  $\lambda$



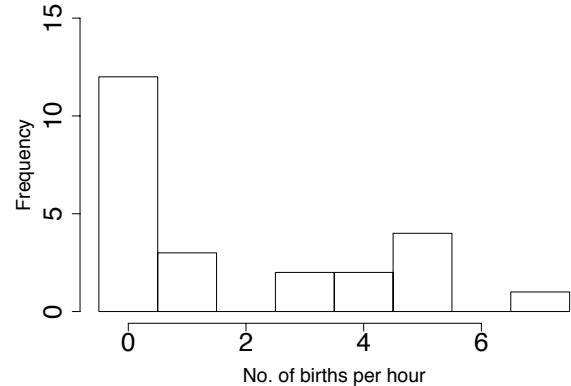
(a) Babyboom data birth times



(b) Histogram of Babyboom birth times



(c) Nonrandom birth times



(d) Histogram of nonrandom birth times

Figure 5.1: Representing the babyboom data set (upper two) and a nonrandom hypothetical collection of birth times (lower two).

**Note** A Poisson random variable can take on any positive integer value. In contrast, the Binomial distribution always has a finite upper limit.

### Example 5.2: Hospital births

Births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability of observing 4 births in a given hour at the hospital?

Let  $X$  = No. of births in a given hour

- (i) Events occur randomly  $\Rightarrow X \sim \text{Po}(1.8)$
- (ii) Mean rate  $\lambda = 1.8$

We can now use the formula to calculate the probability of observing exactly 4 births in a given hour

$$P(X = 4) = e^{-1.8} \frac{1.8^4}{4!} = 0.0723$$

What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

We want  $P(X \geq 2) = P(X = 2) + P(X = 3) + \dots$

i.e. an infinite number of probabilities to calculate

but

$$\begin{aligned} P(X \geq 2) &= P(X = 2) + P(X = 3) + \dots \\ &= 1 - P(X < 2) \\ &= 1 - (P(X = 0) + P(X = 1)) \\ &= 1 - (e^{-1.8} \frac{1.8^0}{0!} + e^{-1.8} \frac{1.8^1}{1!}) \\ &= 1 - (0.16529 + 0.29753) \\ &= 0.537 \end{aligned}$$



### Example 5.3: Disease incidence

Suppose there is a disease, whose average incidence is 2 per million people. What is the probability that a city of 1 million people has at least twice the average incidence?

Twice the average incidence would be 4 cases. We can reasonably suppose the random variable  $X = \#$  cases in 1 million people has Poisson distribution with parameter 2. Then

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - \left( e^{-2} \frac{2^0}{0!} + e^{-2} \frac{2^1}{1!} + e^{-2} \frac{2^2}{2!} + e^{-3} \frac{2^3}{3!} \right) = 0.143.$$

■

### 5.3 The shape of the Poisson distribution

Using the formula we can calculate the probabilities for a specific Poisson distribution and plot the probabilities to observe the shape of the distribution. For example, Figure 5.2 shows 3 different Poisson distributions. We observe that the distributions

- (i). are unimodal;
- (ii). exhibit positive skew (that decreases as  $\lambda$  increases);
- (iii). are centred roughly on  $\lambda$ ;
- (iv). have variance (spread) that increases as  $\lambda$  increases.

### 5.4 Mean and Variance of the Poisson distribution

In general, there is a formula for the mean of a Poisson distribution. There is also a formula for the standard deviation,  $\sigma$ , and variance,  $\sigma^2$ .

If  $X \sim \text{Po}(\lambda)$  then

$$\begin{aligned}\mu &= \lambda \\ \sigma &= \sqrt{\lambda} \\ \sigma^2 &= \lambda\end{aligned}$$

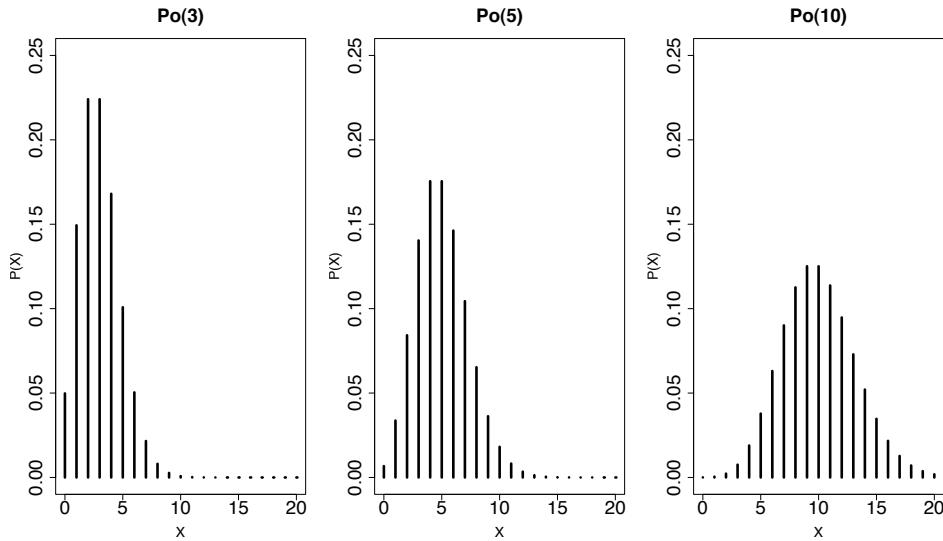


Figure 5.2: Three different Poisson distributions.

## 5.5 Changing the size of the interval

Suppose we know that births in a hospital occur randomly at an average rate of 1.8 births per hour.

What is the probability that we observe 5 births in a given 2 hour interval?

Well, if births occur randomly at a rate of 1.8 births per 1 hour interval  
 Then      births occur randomly at a rate of 3.6 births per 2 hour interval

Let  $Y = \text{No. of births in a 2 hour period}$

Then  $Y \sim \text{Po}(3.6)$

$$P(Y = 5) = e^{-3.6} \frac{3.6^5}{5!} = 0.13768$$

This example illustrates the following rule

If  $X \sim \text{Po}(\lambda)$  on 1 unit interval,  
 then  $Y \sim \text{Po}(k\lambda)$  on  $k$  unit intervals.

## 5.6 Sum of two Poisson variables

Now suppose we know that in hospital A births occur randomly at an average rate of 2.3 births per hour and in hospital B births occur randomly at an average rate of 3.1 births per hour.

What is the probability that we observe 7 births in total from the two hospitals in a given 1 hour period?

To answer this question we can use the following rule

If  $X \sim \text{Po}(\lambda_1)$  on 1 unit interval,  
 and  $Y \sim \text{Po}(\lambda_2)$  on 1 unit interval,  
 then  $X + Y \sim \text{Po}(\lambda_1 + \lambda_2)$  on 1 unit interval.

So if we let  $X = \text{No. of births in a given hour at hospital A}$   
 and  $Y = \text{No. of births in a given hour at hospital B}$

Then  $X \sim \text{Po}(2.3)$ ,  $Y \sim \text{Po}(3.1)$  and  $X + Y \sim \text{Po}(5.4)$

$$\Rightarrow P(X + Y = 7) = e^{-5.4} \frac{5.4^7}{7!} = 0.11999$$

### Example 5.4: Disease Incidence, continued

Suppose disease A occurs with incidence 1.7 per million, and disease B occurs with incidence 2.9 per million. Statistics are compiled, in which these diseases are not distinguished, but simply are all called cases of disease “AB”. What is the probability that a city of 1 million people has at least 6 cases of AB?

If  $Z = \# \text{ cases of AB}$ , then  $P \sim \text{Po}(4.6)$ . Thus,

$$\begin{aligned} P(Z \geq 6) &= 1 - P(Z \leq 5) \\ &= 1 - e^{-4.6} \left( \frac{4.6^0}{0!} + \frac{4.6^1}{1!} + \frac{4.6^2}{2!} + \frac{4.6^3}{3!} + \frac{4.6^4}{4!} + \frac{4.6^5}{5!} \right) \\ &= 0.314. \end{aligned}$$



## 5.7 Fitting a Poisson distribution

Consider the two sequences of birth times we saw in Section 1. Both of these examples consisted of a total of 44 births in 24 hour intervals.

Therefore the mean birth rate for both sequences is  $\frac{44}{24} = 1.8333$

What would be the *expected* counts if birth times were really random i.e. what is the expected histogram for a Poisson random variable with mean rate  $\lambda = 1.8333$ .

Using the Poisson formula we can calculate the probabilities of obtaining each possible value<sup>1</sup>

$x$	0	1	2	3	4	5	$\geq 6$
$P(X = x)$	0.15989	0.29312	0.26869	0.16419	0.07525	0.02759	0.01127

Then if we observe 24 hour intervals we can calculate the expected frequencies as  $24 \times P(X = x)$  for each value of  $x$ .

$x$	0	1	2	3	4	5	$\geq 6$
Expected frequency $24 \times P(X = x)$	3.837	7.035	6.448	3.941	1.806	0.662	0.271

We say we have fitted a Poisson distribution to the data.

This consisted of 3 steps

- (i). Estimating the parameters of the distribution from the data
- (ii). Calculating the probability distribution
- (iii). Multiplying the probability distribution by the number of observations

Once we have fitted a distribution to the data we can compare the expected frequencies to those we actually observed from the real Babyboom dataset. We see that the agreement is quite good.

$x$	0	1	2	3	4	5	$\geq 6$
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	3	8	6	4	3	0	0

---

<sup>1</sup>in practice we group values with low probability into one category.

When we compare the expected frequencies to those observed from the non-random clustered sequence in Section 1 we see that there is much less agreement.

$x$	0	1	2	3	4	5	$\geq 6$
Expected	3.837	7.035	6.448	3.941	1.806	0.662	0.271
Observed	12	3	0	2	2	4	1

In Lecture 9 we will see how we can formally test for a difference between the expected and observed counts. For now it is enough just to know how to fit a distribution.

## 5.8 Using the Poisson to approximate the Binomial

The Binomial and Poisson distributions are both discrete probability distributions. In some circumstances the distributions are very similar. For example, consider the  $\text{Bin}(100, 0.02)$  and  $\text{Po}(2)$  distributions shown in Figure 5.3. Visually these distributions are identical.

In general,

If  $n$  is large (say  $> 50$ ) and  $p$  is small (say  $< 0.1$ ) then a  $\text{Bin}(n, p)$  can be approximated with a  $\text{Po}(\lambda)$  where  $\lambda = np$

### Example 5.5: Counting lefties

Given that 5% of a population are left-handed, use the Poisson distribution to estimate the probability that a random sample of 100 people contains 2 or more left-handed people.

$X$  = No. of left handed people in a sample of 100

$X \sim \text{Bin}(100, 0.05)$

Poisson approximation  $\Rightarrow X \sim \text{Po}(\lambda)$  with  $\lambda = 100 \times 0.05 = 5$

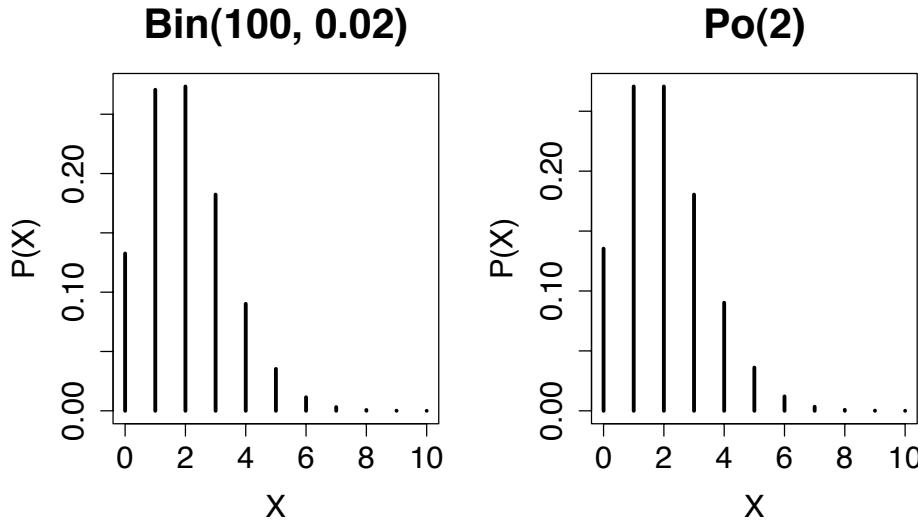


Figure 5.3: A Binomial and Poisson distribution that are very similar.

We want  $P(X \geq 2)$ ?

$$\begin{aligned}
 P(X \geq 2) &= 1 - P(X < 2) \\
 &= 1 - \left( P(X = 0) + P(X = 1) \right) \\
 &\approx 1 - \left( e^{-5} \frac{5^0}{0!} + e^{-5} \frac{5^1}{1!} \right) \\
 &\approx 1 - 0.040428 \\
 &\approx 0.9596
 \end{aligned}$$

If we use the exact Binomial distribution we get the answer 0.9629. ■

The idea of using one distribution to approximate another is widespread throughout statistics and one we will meet again. Why would we use an approximate distribution when we actually know the exact distribution?

- The exact distribution may be hard to work with.
- The exact distribution may have too much detail. There may be some features of the exact distribution that are irrelevant to the questions

we want to answer. By using the approximate distribution, we focus attention on the things we're really concerned with.

For example, consider the Babyboom data, discussed in Example 5.2. We said that “random” birth times should yield numbers of births in each hour that are Poisson distributed. Why? Consider the births between 6 am and 7 am. When we say that the births are random, we probably mean something like this: The times are independent of each other, and have equal chances of happening at any time. Any given one of the 44 births has 24 hours when it could have happened. The probability that it happens during *this* hour is  $p = 1/24 = 0.0417$ . The births between 6 am and 7 am should thus have about the  $\text{Bin}(44, 0.0417)$  distribution. This distribution is about the same as  $\text{Po}(1.83)$ , since  $1.83 = 44 \times 0.0417$ .

### Example 5.6: Drownings in Malta, continued

We now analyse the data on the monthly numbers of drowning incidents in Malta. Under the hypothesis that drownings have nothing to do with each other, and have causes that don't change in time, we would expect the probability the random number  $X$  of drownings occur in a month to have a Poisson distribution? Why is that? We might imagine that there are a large number  $n$  of people in the population, each of whom has an unknown probability  $p$  of drowning in any given month. Then the number of drownings in a month has  $\text{Bin}(n, p)$  distribution. In order to use this model, we need to know what  $n$  and  $p$  are. That is, we need to know the size of the population, which we don't really care about.

On the other hand, the expected (mean) number of monthly drownings is  $np$ , and that can be estimated from the observed mean number of drownings. If we approximate the binomial distribution by  $\text{Po}(\lambda)$ , where  $\lambda = np$ , then we don't have to worry about

We estimate  $\lambda$  as total number of drownings/number of months. The total number of drownings is  $0 \cdot 224 + 1 \cdot 102 + 2 \cdot 23 + 3 \cdot 5 + 4 \cdot 1 = 167$ , so we estimate  $\lambda = 167/355 = 0.47$ . We show the probabilities for the different possible outcomes in the last column of Table 5.2. In the third column we show the *expected* number of months with a given number of drownings, assuming

Table 5.2: Monthly counts of drownings in Malta, with Poisson fit.

No. of drowning deaths per month	Frequency (No. months observed)	Expected frequency Poisson $\lambda = 0.47$	Probability
0	224	221.9	0.625
1	102	104.3	0.294
2	23	24.5	0.069
3	5	3.8	0.011
4	1	0.45	0.001
5+	0	0.04	0.0001

the independence assumption — and hence the Poisson model — is true. This is computed by multiplying the last column by 355. After all, if the probability of no drownings in any given month is 0.625, and we have 355 months of observations, we expect  $0.625 \cdot 355$  months with 0 drownings.

We see that the observations (in the second column) are pretty close to the predictions of the Poisson model (in the third column), so the data do not give us strong evidence to reject the neutral assumption, that drownings are independent of one another, and have a constant rate in time. In Lecture 9 we will describe one way of testing this hypothesis formally. ■

### Example 5.7: Swine flu vaccination

In 1976, fear of an impending swine flu pandemic led to a mass vaccination campaign in the US. The pandemic never materialised, but there were concerns that the vaccination may have led to an increase in a rare and serious neurological disease, Guillain-Barré Syndrome (GBS). It was difficult to determine whether the vaccine was really at fault, since GBS may arise spontaneously — about 1 person in 100,000 develops GBS in a given year — and the number of cases was small.

Consider the following data from the US state of Michigan: Out of 9 million residents, about 2.3 million were vaccinated. Of

those, 48 developed GBS between July 1976 and June 1977. We might have expected

$$2.3 \text{ million} \times 10^{-5} \text{ cases/person-year} = 23 \text{ cases.}$$

How likely is it that, purely by chance, this population would have experienced 48 cases in a single year? If  $Y$  is the number of cases, it would then have Poisson distribution with parameter 23, so that

$$P(Y \geq 48) = 1 - \sum_{i=0}^{47} e^{-23} \frac{23^i}{i!} = 3.5 \times 10^{-6}.$$

So, such an extreme number of cases is likely to happen less than 1 year in 100,000. Does this prove that the vaccine caused GBS?

The people who had the vaccine are people who **chose** to be vaccinated. They may differ from the rest of the population in multiple ways in addition to the elementary fact of having been vaccinated, and some of those ways may have predisposed them to GBS. What can we do? The paper [BH84] takes the following approach: If the vaccine were not the cause of the GBS cases, we would expect no connection between the timing of the vaccine and the onset of GBS. In fact, though, there seemed to be a particularly large number of cases in the six weeks following vaccination. Can we say that this was more than could reasonably be expected by chance?

The data are given in Table 5.3. Each of the 40 GBS cases was assigned a time, which is the number of weeks after vaccination when the disease was diagnosed. (Thus “week 1” is a different calendar week for each subject.) If the cases are evenly distributed, the number in a given week should be Poisson distributed with parameter  $40/30 = 1.33$ . Using this parameter, we compute the probabilities of 0, 1, 2, … cases in a week, which we give in row 3 of Table 5.3. Multiplying these numbers by 30 gives the expected frequencies in row 4 of the table. It is clear that the observed and expected frequencies are very different. One way of seeing this is to consider the standard deviation. The Poisson distribution has  $\text{SD } \sqrt{1.33} = 1.15$  (as discussed in section 5.4,

while the data have SD

$$s = \sqrt{\frac{1}{30-1} \left( 16 \cdot (0 - 1.33)^2 + 7 \cdot (1 - 1.33)^2 + 3 \cdot (2 - 1.33)^2 + 2 \cdot (4 - 1.33)^2 + 1 \cdot (9 - 1.33)^2 + 1 \cdot (10 - 1.33)^2 \right)} = 2.48.$$

Table 5.3: Cases of GBS, by weeks after vaccination

# cases per week	0	1	2	3	4	5	6+
observed frequency	16	7	3	0	2	0	2
probability	0.264	0.352	0.234	0.104	0.034	0.009	0.003
expected frequency	7.9	10.6	7.0	3.1	1.0	0.3	0.1

■

## 5.9 Derivation of the Poisson distribution (non-examinable)

This section is not officially part of the course, but is optional, for those who are interested in more mathematical detail. Where does the formula in section 5.2 come from?

Think of the Poisson distribution as in section 5.8, as an approximation to a binomial distribution. Let  $X$  be the (random) number of successes in a collection of independent random trials, where the expected number of successes is  $\lambda$ . This will, of course, depend on the number of trials, but we show that when the number of trials (call it  $n$ ) gets large, the **exact** number of trials doesn't matter. In mathematical language, we say that the probability *converges* to a *limit* as  $n$  goes to infinity. But how large is "large"? We would like to know how good the approximation is, for real values of  $n$ , of the sort that we are interested in.

Let  $X_n$  be the random number of successes in  $n$  independent trials, where the probability of each success is  $\lambda/n$ . Thus, the probability of success goes down as the number of trials goes up, and expected number of successes is always the same  $\lambda$ . Then

$$P\{X_n = x\} = {}^n C_x \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}.$$

Now, those of you who have learned some calculus at A-levels may remember the Taylor series for  $e^z$ :

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots$$

In particular, for small  $z$  we have  $e^{-z} \approx 1 - z$ , and the difference (or “error” in the approximation) is no bigger than  $z^2/2$ . The key idea is that if  $z$  is very small (as it is when  $z = \lambda/n$ , and  $n$  is large), then  $z^2$  is a lot smaller than  $z$ .

Using a bit of algebra, we have

$$\begin{aligned} P\{X_n = x\} &= {}^n C_x \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^n \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!} \left(\frac{\lambda^x}{n^x}\right) \left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^n \\ &= \frac{\lambda^x}{x!} \frac{(1 - \frac{1}{n})\cdots(1 - \frac{x-1}{n})}{\left(1 - \frac{\lambda}{n}\right)^x} \left(1 - \frac{\lambda}{n}\right)^n. \end{aligned}$$

Now, if we’re not concerned about the size of the error, we can simply say that  $n$  is much bigger than  $\lambda$  or  $x$  (because we’re thinking of a fixed  $\lambda$  and  $x$ , and  $n$  getting large). So we have the approximations

$$\begin{aligned} \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) &\approx 1; \\ \left(1 - \frac{\lambda}{n}\right)^x &\approx 1; \\ \left(1 - \frac{\lambda}{n}\right)^n &\approx \left(e^{-\lambda/n}\right)^n = e^{-\lambda}. \end{aligned}$$

Thus

$$P\{X_n = x\} \approx \frac{\lambda^x}{x!} e^{-\lambda}.$$

### 5.9.1 Error bounds (very mathematical)

In the long run,  $X_n$  has a distribution very close to the Poisson distribution defined in section 5.2. But how long is “the long run”? Do we need 10 trials? 1000? a billion?

If you just want the answer, it’s approximately this: The error that you’ll make by taking the Poisson distribution instead of the binomial is no more

than about  $1.6\lambda^2/n^{3/2}$ . In Example 5.5, where  $n = 100$  and  $\lambda = 5$ , this says the error won't be bigger than about 0.04, which is useful information, although in reality the maximum error is about 10 times smaller than this. On the other hand, if  $n = 400,000$  (about the population of Malta), and  $\lambda = 0.47$ , then the error will be only about  $10^{-8}$ .

Let's assume that  $n$  is at least  $4\lambda^2$ , so  $\lambda < \sqrt{n}/2$ . Define the *approximation error* to be

$$\epsilon := \max |P\{X_n = x\} - P\{X = x\}|.$$

(The bars  $|\cdot|$  mean that we're only interested in how big the difference is, not whether it's positive or negative.) Then

$$\begin{aligned} P\{X_n = x\} - P\{X = x\} &= \frac{\lambda^x}{x!} \left( \frac{(1) \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right)}{\left(1 - \frac{\lambda}{n}\right)^x} \left(1 - \frac{\lambda}{n}\right)^n - \frac{\lambda^x}{x!} e^{-\lambda} \right) \\ &= \frac{\lambda^x}{x!} e^{-\lambda} \left( (1) \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{-x} \left(\frac{1-\lambda/n}{e^{-\lambda/n}}\right)^n - 1 \right) \end{aligned}$$

If  $x$  is bigger than  $\sqrt{n}$ , then  $P\{X = x\}$  and  $P\{X_n = x\}$  are both tiny; we won't go into the details here, but we will consider only  $x$  that are smaller than this. Now we have to do some careful approximation. Basic algebra tells us that if  $a$  and  $b$  are positive,

$$(1-a)(1-b) = 1 - (a+b) + ab > 1 - (a+b).$$

We can extend this to  $(1-a)(1-b)(1-c) > (1-(a+b))(1-c) > 1-(a+b+c)$ . And so, finally, if  $a, b, c, \dots$  are all positive, then

$$1 > (1-a)(1-b)(1-c) \cdots (1-z) > 1 - (a+b+c+\cdots+z).$$

Thus

$$1 - \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x-1}{n}\right) > 1 - \sum_{k=0}^{x-1} \frac{k}{n} > 1 - \frac{x^2}{2n},$$

and

$$1 > \left(1 - \frac{\lambda}{n}\right)^x > 1 - \frac{\lambda x}{n}.$$

Again applying some calculus, we turn this into

$$1 < \left(1 - \frac{\lambda}{n}\right)^{-x} < 1 + \frac{\lambda x}{n - \lambda x}.$$

We also know that

$$1 - \frac{\lambda}{n} < e^{-\lambda/n} < 1 - \frac{\lambda}{n} + \frac{\lambda^2}{2n^2},$$

which means that

$$1 - \frac{\lambda^2}{2(n^2 - n\lambda)} < \frac{1 - \lambda/n}{e^{-\lambda/n}} < 1,$$

and

$$1 - \frac{\lambda^2}{2(n - \lambda)} < \left(1 - \frac{\lambda^2}{2(n^2 - n\lambda)}\right)^n < \left(\frac{1 - \lambda/n}{e^{-\lambda/n}}\right)^n < 1.$$

Now we put together all the overestimates on one side, and all the underestimates on the other.

$$\frac{\lambda^x}{x!} e^{-\lambda} \left( -\frac{\lambda^2}{2(n - \lambda)} - \frac{\lambda x}{n} \right) \leq P\{X_n = x\} - P\{X = x\} \leq \frac{\lambda^x}{x!} e^{-\lambda} \left( \frac{\lambda x}{n - \lambda x} \right).$$

So, finally, as long as  $n \geq 4\lambda^2$ , we get

$$\epsilon \leq \max \frac{\lambda^{x+1}}{x!} e^{-\lambda} \left( \frac{\lambda}{n} + \frac{x}{n} + \frac{x}{n(1 - x/2\sqrt{n})} \right).$$

We need to find the maximum over all possible  $x$ . If  $x < \sqrt{n}$  then this becomes

$$\epsilon \leq \max \frac{1}{n} \frac{\lambda^{x+1}}{x!} e^{-\lambda} (\lambda + 3x) \leq \frac{4\lambda_*^2}{n\sqrt{2\pi n}},$$

(by a formula known as “Stirling’s formula”), where  $\lambda_* = \max\{\lambda, 1\}$ .

## Lecture 6

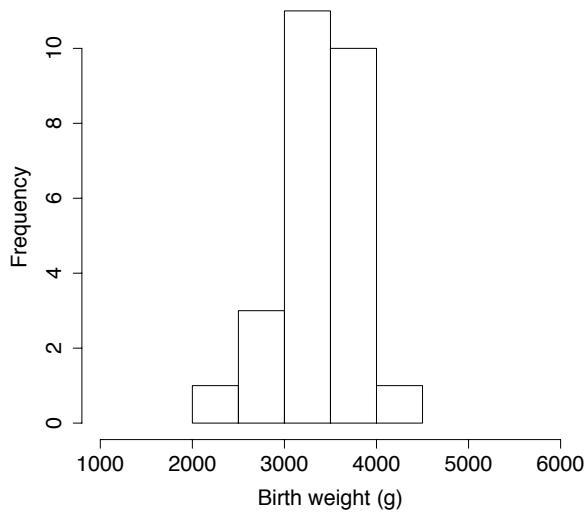
# The Normal Distribution

### 6.1 Introduction

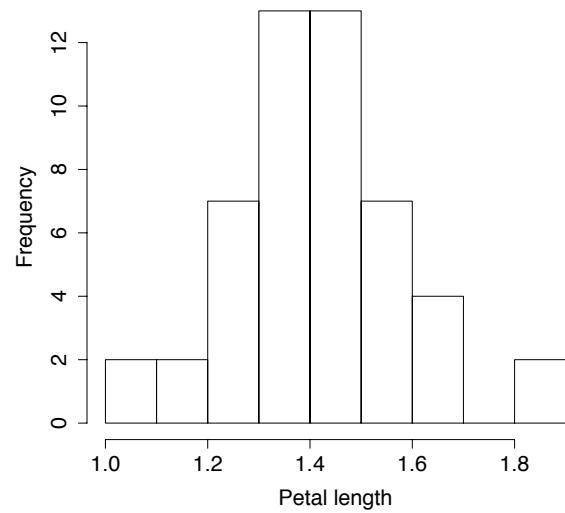
In previous lectures we have considered discrete datasets and discrete probability distributions. In practice many datasets that we collect from experiments consist of continuous measurements. For example, there are the weights of newborns in the babyboom data set (Table 1.2). The plots in Figure 6.1 show histograms of real datasets consisting of continuous measurements. From such samples of continuous data we might want to test whether the data is consistent with a specific population mean value or whether there is a significant difference between 2 groups of data. To answer these question we need a probability model for the data. Of course, there are many different possible distributions that quantities could have. It is therefore a startling fact that many different quantities that we are commonly interested in — heights, weights, scores on intelligence tests, serum potassium levels of different patients, measurement errors of distance to the nearest star — all have distributions which are close to one particular shape. This shape is called the **Normal** or **Gaussian**<sup>1</sup> family of distributions.

---

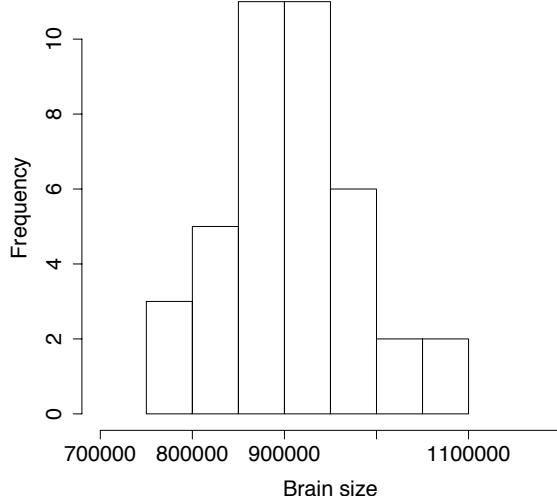
<sup>1</sup>Named for German mathematician Carl Friedrich Gauss, who first worked out the formula for these distributions, and used them to estimate the errors in astronomical computations. Until the introduction of the euro, Gauss's picture — and the Gaussian curve — were on the German 10 mark banknote.



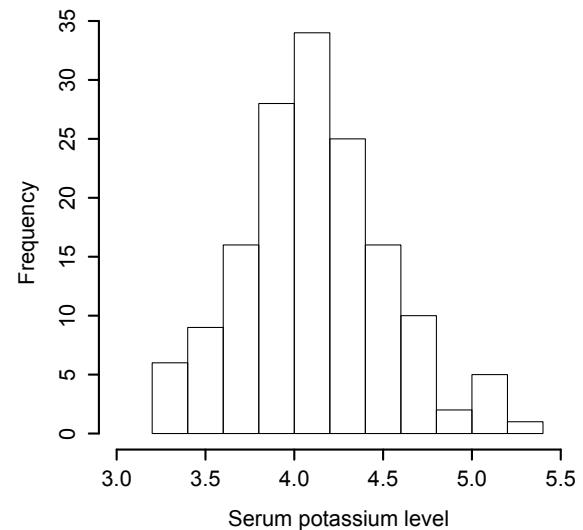
(a) Babyboom birthweights



(b) Petal measurements in a species of flower



(c) Brain sizes of 40 Psychology students



(d) Serum potassium measurements from 152 healthy volunteers

Figure 6.1: Histograms of some continuous data.

## 6.2 Continuous probability distributions

When we considered the Binomial and Poisson distributions we saw that the probability distributions were characterized by a formula for the probability of each possible discrete value. All of the probabilities together sum up to 1. We can visualize the density by plotting the probabilities against the discrete values (Figure 6.2). For continuous data we don't have equally spaced discrete values so instead we use a curve or function that describes the probability *density* over the range of the distribution (Figure 6.3). The curve is chosen so that the area under the curve is equal to 1. If we observe a sample of data from such a distribution we should see that the values occur in regions where the density is highest.

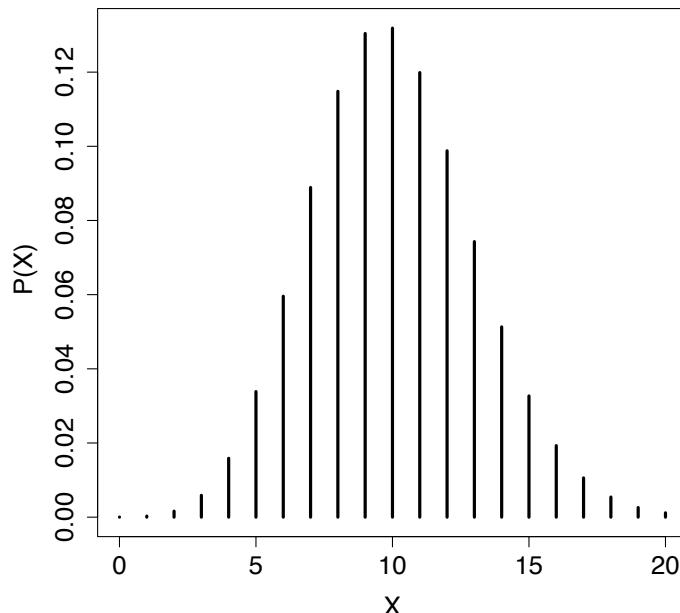


Figure 6.2: A discrete probability distribution

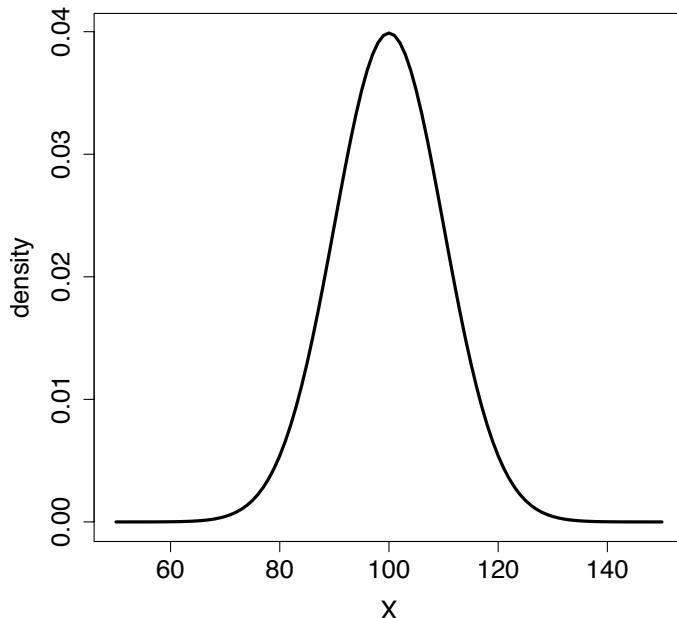


Figure 6.3: A continuous probability distribution

### 6.3 What is the Normal Distribution?

There will be many, many possible probability density functions over a continuous range of values. The Normal distribution describes a special class of such distributions that are symmetric and can be described by the distribution mean  $\mu$  and the standard deviation  $\sigma$  (or variance  $\sigma^2$ ). 4 different Normal distributions are shown in Figure 6.4 together with the values of  $\mu$  and  $\sigma$ . These plots illustrate how changing the values of  $\mu$  and  $\sigma$  alter the positions and shapes of the distributions.

If  $X$  is Normally distributed with mean  $\mu$  and standard deviation  $\sigma$ , we write

$$X \sim N(\mu, \sigma^2)$$

$\mu$  and  $\sigma$  are the **parameters** of the distribution.

The probability density of the Normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-(x-\mu)^2/2\sigma^2}$$

For the purposes of this course we do not need to use this expression. It is included here for future reference.

---

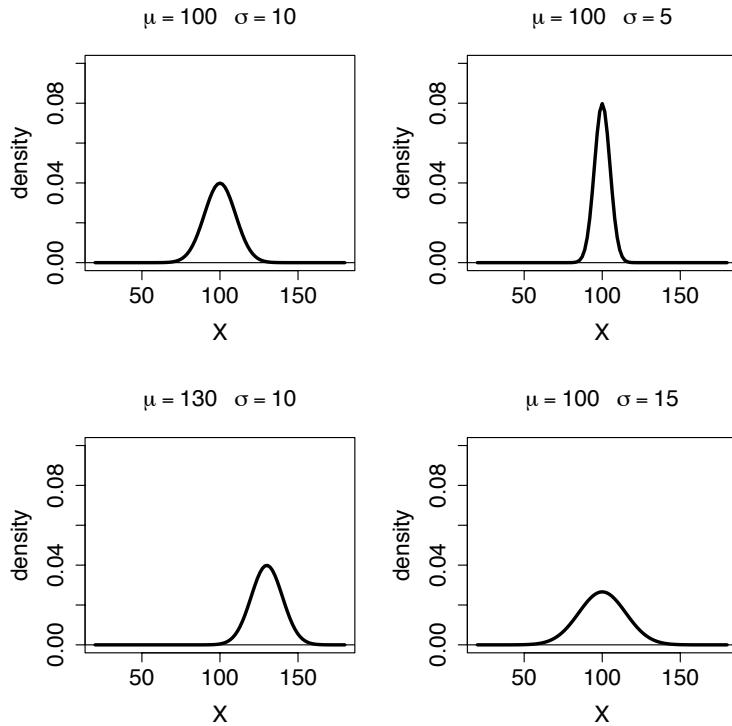


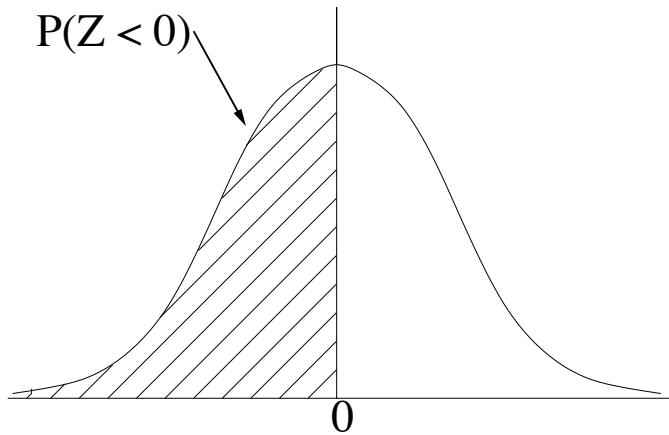
Figure 6.4: 4 different Normal distributions

## 6.4 Using the Normal table

For a discrete probability distribution we calculate the probability of being less than some value  $z$ , i.e.  $P(Z < z)$ , by simply summing up the probabilities of the values less than  $z$ .

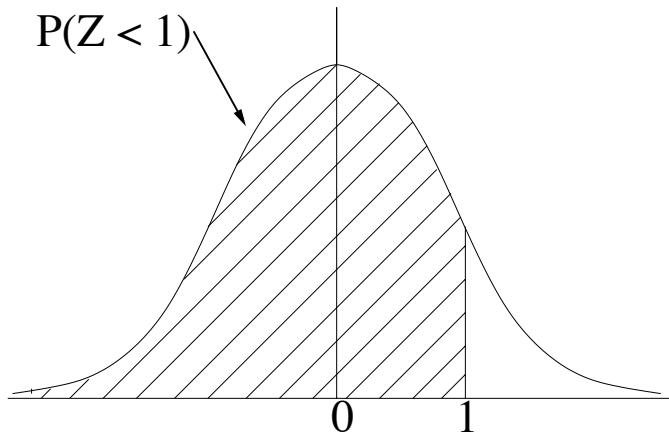
For a continuous probability distribution we calculate the probability of being less than some value  $z$ , i.e.  $P(Z < z)$ , by calculating the area under the curve to the left of  $z$ .

For example, suppose  $Z \sim N(0, 1)$  and we want to calculate  $P(Z < 0)$ ?



For this example we can calculate the required area as we know the distribution is symmetric and the total area under the curve is equal to 1, i.e.  $P(Z < 0) = 0.5$ .

What about  $P(Z < 1.0)$ ?



Calculating this area is not easy<sup>2</sup> and so we use probability tables. Probabil-

---

<sup>2</sup>For those Mathematicians who recognize this area as a definite integral and try to do the integral by hand please note that the integral *cannot* be evaluated analytically

ity tables are tables of probabilities that have been calculated on a computer. All we have to do is identify the right probability in the table and copy it down! Obviously it is impossible to tabulate all possible probabilities for all possible Normal distributions so only one special Normal distribution,  $N(0, 1)$ , has been tabulated.

The  $N(0, 1)$  distribution is called the **standard Normal distribution**.

The tables allow us to read off probabilities of the form  $P(Z < z)$ . Most of the table in the formula book has been reproduced in Table 6.1. From this table we can identify that  $P(Z < 1.0) = 0.8413$  (this probability has been highlighted with a box).

$z$	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	5040	5080	5120	5160	5199	5239	5279	5319	5359
0.1	0.5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
0.2	0.5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
0.3	0.6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
0.4	0.6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
0.5	0.6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
0.6	0.7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
0.7	0.7580	7611	7642	7673	7704	7734	7764	7794	7823	7852
0.8	0.7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
0.9	0.8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1.0	0.8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1.1	0.8643	8665	8686	8708	8729	8749	8770	8790	8810	8830

Table 6.1:  $N(0, 1)$  probability table

Once we know how to read tables we can calculate other probabilities.

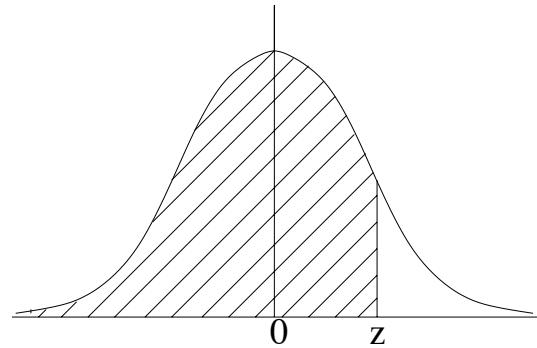
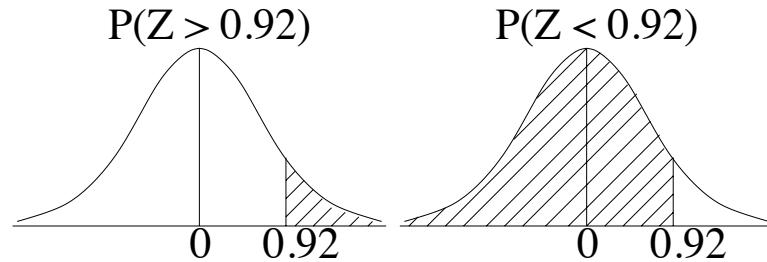
### Example 6.1: Upper tails

The table gives  $P(Z < z)$ . Suppose we want  $P(Z > 0.92)$ .

We know that  $P(Z > 0.92) = 1 - P(Z < 0.92)$  and we can calculate  $P(Z < 0.92)$  from the tables.

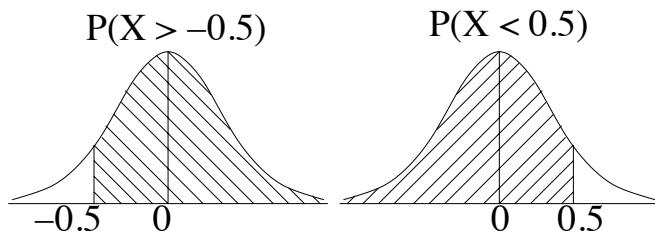
Thus,  $P(Z > 0.92) = 1 - 0.8212 = 0.1788$ .





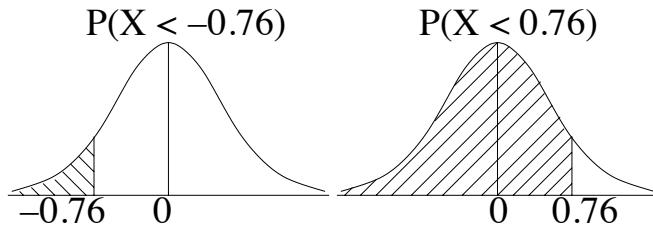
### Example 6.2: Negative $z$

The table only includes positive values of  $z$ . What about negative values? Compute  $P(Z > -0.5)$ .



The Normal distribution is symmetric so we know that  $P(Z > -0.5) = P(Z < 0.5) = 0.6915$

We can use the symmetry of the Normal distribution to calculate

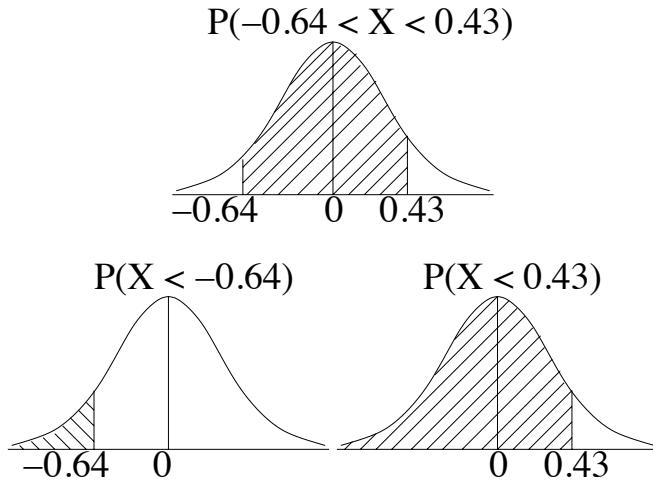


$$\begin{aligned}
 P(Z < -0.76) &= P(Z > 0.76) \\
 &= 1 - P(Z < 0.76) \\
 &= 1 - 0.7764 \\
 &= 0.2236.
 \end{aligned}$$

■

**Example 6.3: Intervals**

How do we compute  $P(-0.64 < Z < 0.43)$ ?



We can calculate this using

$$\begin{aligned}
 P(-0.64 < Z < 0.43) &= P(Z < 0.43) - P(Z < -0.64) \\
 &= 0.6664 - (1 - 0.7389) \\
 &= 0.4053.
 \end{aligned}$$

■

### Example 6.4: Interpolation

How would we compute  $P(Z < 0.567)$ ?

From tables we know that  $P(Z < 0.56) = 0.7123$  and  $P(Z < 0.57) = 0.7157$

To calculate  $P(Z < 0.567)$  we *interpolate* between these two values

$$P(Z < 0.567) = 0.3 \times 0.7123 + 0.7 \times 0.7157 = 0.7146$$

■

## 6.5 Standardisation

All of the probabilities above were calculated for the standard Normal distribution  $N(0, 1)$ . If we want to calculate probabilities from different Normal distributions we convert the probability to one involving the standard Normal distribution. This process is called **standardisation**.

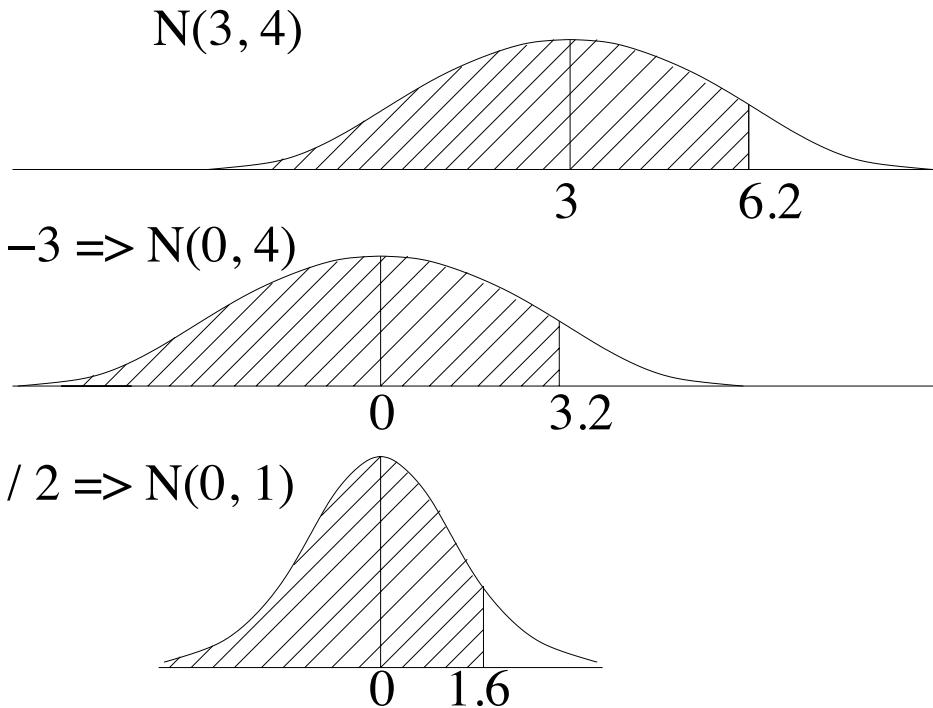
Suppose  $X \sim N(3, 4)$  and we want to calculate  $P(X < 6.2)$ . We convert this probability to one involving the  $N(0, 1)$  distribution by

- (i). Subtracting the mean  $\mu$
- (ii). Dividing by the standard deviation  $\sigma$

Subtracting the mean re-centers the distribution on zero. Dividing by the standard deviation re-scales the distribution so it has standard deviation 1. If we also transform the boundary point of the area we wish to calculate we obtain the equivalent boundary point for the  $N(0, 1)$  distribution. This process is illustrated in the figure below. In this example,  $P(X < 6.2) = P(Z < 1.6) = 0.9452$  where  $Z \sim N(0, 1)$

This process can be described by the following rule

$$\text{If } X \sim N(\mu, \sigma^2) \text{ and } Z = \frac{X-\mu}{\sigma} \text{ then } Z \sim N(0, 1)$$



### Example 6.5: Birth weights

Suppose we know that the birth weight of babies is Normally distributed with mean 3500g and standard deviation 500g. What is the probability that a baby is born that weighs less than 3100g?

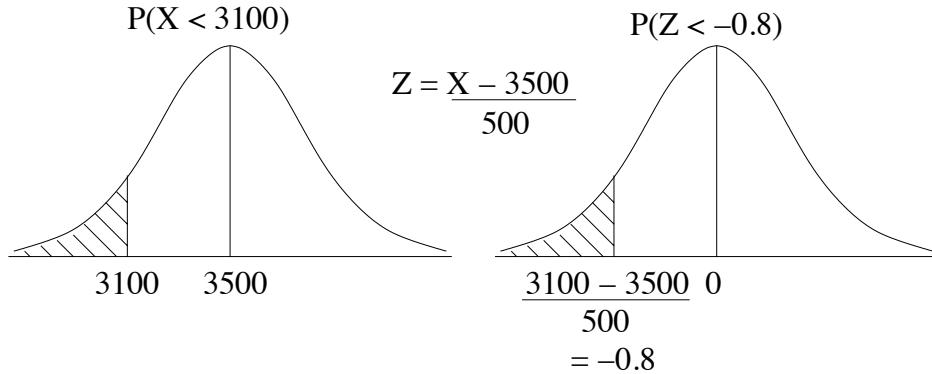
That is  $X \sim N(3500, 500^2)$  and we want to calculate  $P(X < 3100)$ ?

We can calculate the probability through the process of standardization.

Drawing a rough diagram of the process can help you to avoid any confusion about which probability (area) you are trying to calculate.

$$X \sim N(3500, 500^2)$$

$$Z \sim N(0, 1)$$



$$\begin{aligned} P(X < 3100) &= P\left(\frac{X - 3500}{500} < \frac{3100 - 3500}{500}\right) \\ &= P(Z < -0.8) \quad \text{where } Z \sim N(0, 1) \\ &= 1 - P(Z < 0.8) \\ &= 1 - 0.7881 \\ &= 0.2119 \end{aligned}$$

■

## 6.6 Linear combinations of Normal random variables

Suppose two rats A and B have been trained to navigate a large maze. The time it takes rat A is normally distributed with mean 80 seconds and standard deviation 10 seconds. The time it takes rat B is normally distributed with mean 78 seconds and standard deviation 13 seconds. On any given day what is the probability that rat A runs the maze faster than rat B?

$$\begin{array}{ll} X = \text{Time of run for rat A} & X \sim N(80, 10^2) \\ Y = \text{Time of run for rat B} & Y \sim N(78, 13^2) \end{array}$$

Let  $D = X - Y$  be the difference in times of rats A and B

If rat A is faster than rat B then  $D < 0$  so we want  $P(D < 0)$ ?

To calculate this probability we need to know the distribution of  $D$ . To do this we use the following rule

If  $X$  and  $Y$  are two independent normal variable such that

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2)$$

$$\text{then } X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$$

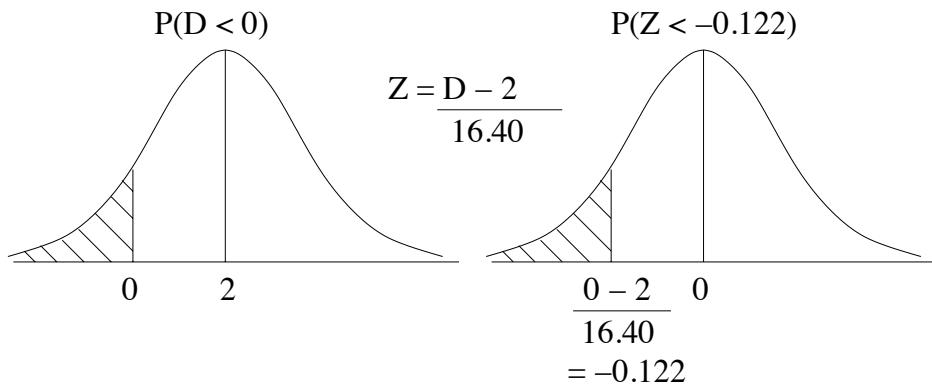
In this example,

$$D = X - Y \sim N(80 - 78, 10^2 + 13^2) = N(2, 269)$$

We can now calculate this probability through standardisation

$$D \sim N(2, 269)$$

$$Z \sim N(0, 1)$$



$$\begin{aligned}
 P(D < 0) &= P\left(\frac{D - 2}{\sqrt{269}} < \frac{0 - 2}{\sqrt{269}}\right) = P(Z < -0.122) \quad \text{where } Z \sim N(0, 1) \\
 &= 1 - (0.8 \times 0.5478 + 0.2 \times 0.5517) \\
 &= 0.45142
 \end{aligned}$$

Other rules that are often used are

If  $X$  and  $Y$  are two independent normal variable such that

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2)$$

then

$$\begin{aligned}
 X + Y &\sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \\
 aX &\sim N(a\mu_1, a^2\sigma_1^2) \\
 aX + bY &\sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)
 \end{aligned}$$

### Example 6.6: Maze-running times

Suppose two rats A and B have been trained to navigate a large maze. The time it takes rat A is normally distributed with mean 80 seconds and standard deviation 10 seconds. The time it takes rat B is normally distributed with mean 78 seconds and standard deviation 13 seconds. On any given day what is the probability that the average time the rats take to run the maze is greater than 82 seconds?

$$\begin{aligned}
 X &= \text{Time of run for rat A} & X &\sim N(80, 10^2) \\
 Y &= \text{Time of run for rat B} & Y &\sim N(78, 13^2)
 \end{aligned}$$

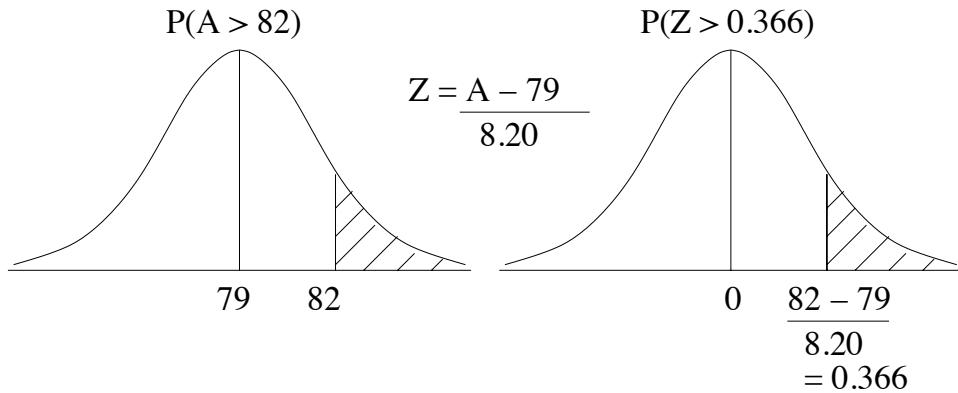
Let  $A = \frac{X+Y}{2} = \frac{1}{2}X + \frac{1}{2}Y$  be the average time of rats A and B

$$\text{Then } A \sim N\left(\frac{1}{2}80 + \frac{1}{2}78, (\frac{1}{2})^210^2 + (\frac{1}{2})^213^2\right) = N(79, 67.25)$$

We want  $P(A > 82)$

$$A \sim N(79, 67.25)$$

$$Z \sim N(0, 1)$$



$$\begin{aligned}
 P(A > 82) &= P\left(\frac{A - 79}{\sqrt{67.25}} < \frac{82 - 79}{\sqrt{67.25}}\right) = P(Z > 0.366) \quad \text{where } Z \sim N(0, 1) \\
 &= 1 - (0.4 \times 0.6406 + 0.6 \times 0.6443) \\
 &= 0.35718
 \end{aligned}$$

■

## 6.7 Using the Normal tables backwards

### Example 6.7: Exam scores

The marks of 500 candidates in an examination are normally distributed with a mean of 45 marks and a standard deviation of 20 marks.

If 20% of candidates obtain a distinction by scoring  $x$  marks or more, estimate the value of  $x$ .

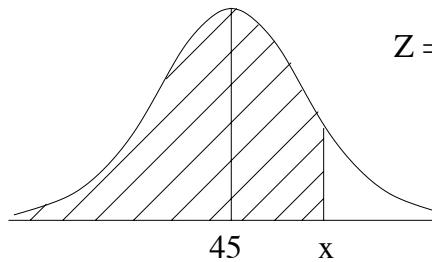
We have  $X \sim N(45, 20^2)$  and we want  $x$  such that  $P(X > x) = 0.2$

$$\Rightarrow P(X < x) = 0.8$$

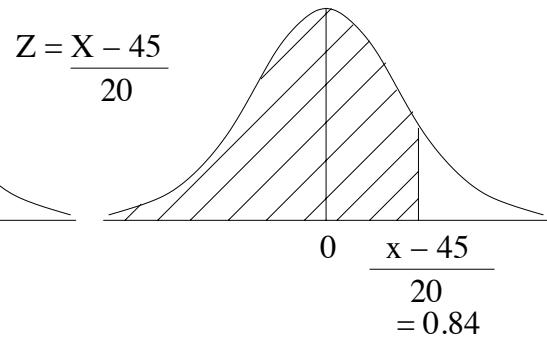
$$X \sim N(45, 400)$$

$$Z \sim N(0, 1)$$

$$P(X < x) = 0.8$$



$$P(Z < 0.84) = 0.8$$



Standardising this probability we get

$$\begin{aligned} P\left(\frac{X - 45}{20} < \frac{x - 45}{20}\right) &= 0.8 \\ \Rightarrow P\left(Z < \frac{x - 45}{20}\right) &= 0.8 \end{aligned}$$

From the tables we know that  $P(Z < 0.84) \approx 0.8$  so

$$\begin{aligned} \frac{x - 45}{20} &\approx 0.84 \\ \Rightarrow x &\approx 45 + 20 \times 0.84 = 61.8 \end{aligned}$$

■

## 6.8 The Normal approximation to the Binomial

Under certain conditions we can use the Normal distribution to approximate the Binomial distribution. This can be very useful when we need to sum up

a large number of Binomial probabilities to calculate the probability that we want.

For example, Figure 6.5 compares a  $\text{Bin}(300, 0.5)$  and a  $\text{N}(150, 75)$  which both have the same mean and variance. The figure shows that the distributions are very similar.

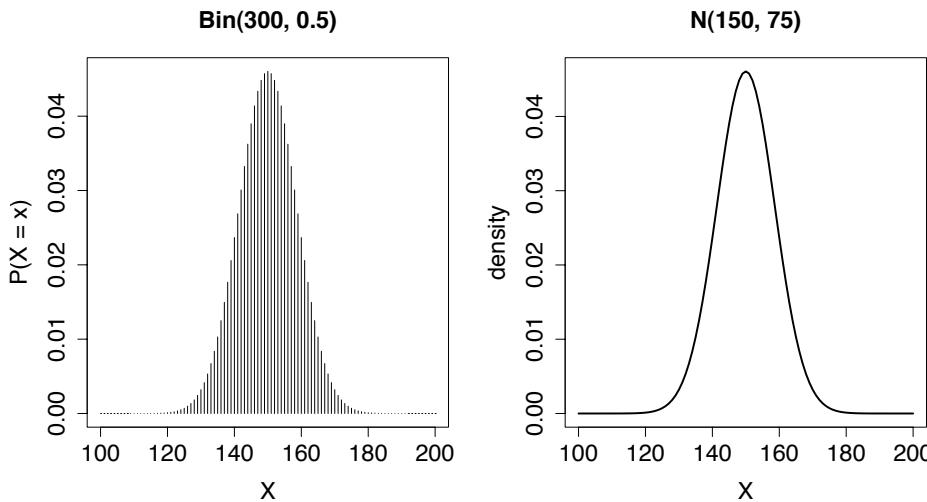


Figure 6.5: Comparison of a  $\text{Bin}(300, 0.5)$  and a  $\text{N}(150, 75)$  distribution

In general,

If  $X \sim \text{Bin}(n, p)$  then

$$\begin{aligned}\mu &= np \\ \sigma^2 &= npq \quad \text{where } q = 1 - p\end{aligned}$$

For large  $n$  and  $p$  not too small or too large

$$X \sim \text{N}(np, npq)$$

$$n > 10 \text{ and } p \approx \frac{1}{2} \text{ OR } n > 30 \text{ and } p \text{ moving away from } \frac{1}{2}$$

### 6.8.1 Continuity correction

Suppose  $X \sim \text{Bin}(12, 0.5)$  what is  $P(4 \leq X \leq 7)$ ?

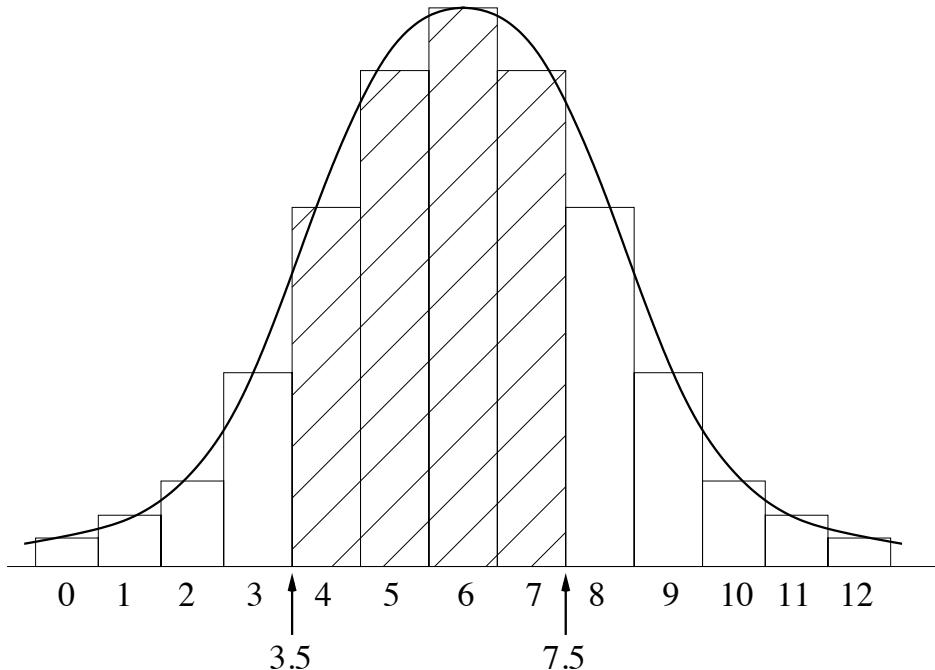
For this distribution we have

$$\begin{aligned}\mu &= np = 6 \\ \sigma^2 &= npq = 3\end{aligned}$$

So we can use a  $N(6, 3)$  distribution as an approximation.

Unfortunately, it's not quite so simple. We have to take into account the fact that we are using a *continuous* distribution to approximate a *discrete* distribution. This is done using a **continuity correction**. The continuity correction appropriate for this example is illustrated in the figure below

In this example,  $P(4 \leq X \leq 7)$  transforms to  $P(3.5 < X < 7.5)$



$$\begin{aligned}P(3.5 < X < 7.5) &= P\left(\frac{3.5 - 6}{\sqrt{3}} < \frac{X - 6}{\sqrt{3}} < \frac{7.5 - 6}{\sqrt{3}}\right) \\ &= P(-1.443 < Z < 0.866) \quad \text{where } Z \sim N(0, 1) \\ &= 0.732\end{aligned}$$

The exact answer is 0.733 so in this case the approximation is very good.

## 6.9 The Normal approximation to the Poisson

We can also use the Normal distribution to approximate a Poisson distribution under certain conditions.

In general,

If  $X \sim \text{Po}(\lambda)$  then

$$\begin{aligned}\mu &= \lambda \\ \sigma^2 &= \lambda\end{aligned}$$

For large  $\lambda$  (say  $\lambda > 20$ )

$$X \sim N(\lambda, \lambda)$$

### Example 6.8: Radioactive emission

A radioactive source emits particles at an average rate of 25 particles per second. What is the probability that in 1 second the count is less than 28 particles?

$$X = \text{No. of particles emitted in 1 second} \quad X \sim \text{Po}(25)$$

So, we can use a  $N(25, 25)$  as an approximate distribution.

Again, we need to make a continuity correction

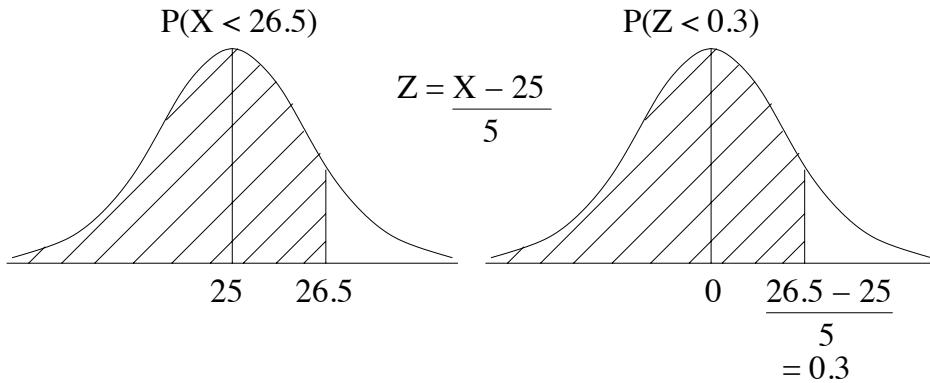
So  $P(X < 27)$  transforms to  $P(X < 26.5)$

$$\begin{aligned}P(X < 26.5) &= P\left(\frac{X - 25}{5} < \frac{26.5 - 25}{5}\right) \\ &= P(Z < 0.3) \quad \text{where } Z \sim N(0, 1) \\ &= 0.6179\end{aligned}$$



$$X \sim N(25, 25)$$

$$Z \sim N(0, 1)$$



### Example 6.9: ESP Experiment

*This example is adapted from [FPP98].*

In the 1970s, the psychologist Charles Tart tried to test whether people might have the power to see into the future. His “Aquarius machine” was a device that would flash four different lights in random orders. Subjects would press buttons to predict which of the 4 lights will come on next.

15 different subjects each ran a trial of 500 guesses, so 1500 guesses in total. They produced 2006 correct guesses and 5494 incorrect. What should we conclude?

We might begin by hypothesising that without any power to predict the future, a subject has just a  $1/4$  chance of guessing right each time, independent of any other outcomes. Thus, the number of correct guesses  $X$  has  $\text{Bin}(7500, 1/4)$  distribution. This has mean  $\mu = 7500/4 = 1875$ , and standard deviation  $\sqrt{7500 \times 0.25 \times 0.75} = 37.5$ . The result is thus above the mean: There were more correct guesses than would be expected. Might we plausibly say that the difference from the expectation is just chance variation?

We want to know how likely a result this extreme would be if  $X$  really has this binomial distribution. We could compute this

directly from the binomial distribution as

$$\begin{aligned} P(X \geq 2006) &= \sum_{x=2006}^{7500} \binom{7500}{x} (0.25)^x (0.75)^{7500-x} \\ &= \binom{7500}{2006} (0.25)^{2006} (0.75)^{5494} \\ &\quad + \binom{7500}{2007} (0.25)^{2007} (0.75)^{5493} + \dots \\ &\quad + \binom{7500}{7500} (0.25)^{7500} (0.75)^0. \end{aligned}$$

This is not only a lot of work, it is also not very illuminating. More useful is to treat  $X$  as a continuous variable that is approximately normal.

We sketch the relevant normal curve in Figure 6.7. This is the normal distribution with mean 1875 and SD 37.5. Because of the continuity correction, the probability we are looking for is  $P(X > 2005.5)$ . We convert  $x = 2005.5$  into standard units:

$$z = \frac{x - \mu}{\sigma} = \frac{2005.5 - 1875}{37.5} = 3.48.$$

(Note that with such a large SD, the continuity correction makes hardly any difference.) We have then  $P(X > 2005.5) \approx P(Z > 3.48)$ , where  $Z$  has standard normal distribution. Since most of the probability of the standard normal distribution is between  $-2$  and  $2$ , and nearly all between  $-3$  and  $3$ , we know this is a small probability. The relevant piece of the normal table is given in Figure 6.6. (Notice that the table has become less refined for  $z > 2$ , giving only one place after the decimal point in  $z$ .) From the table we see that

$$P(X > 2005.5) = P(Z > 3.48) = 1 - P(Z < 3.48),$$

which is between 0.0002 and 0.0003. (Using a more refined table, we would see that  $P(Z > 3.48) = 0.000250\dots$ ) This may be compared to the exact binomial probability 0.000274.

■

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
2.0	0.9772	9821	9861	9893	9918	9938	9953	9965	9974	9981
3.0	0.9987	9990	9993	9995	9997	9998	9998	9999	9999	9999

Figure 6.6: Normal table used to compute tail probability for Aquarius experiment.

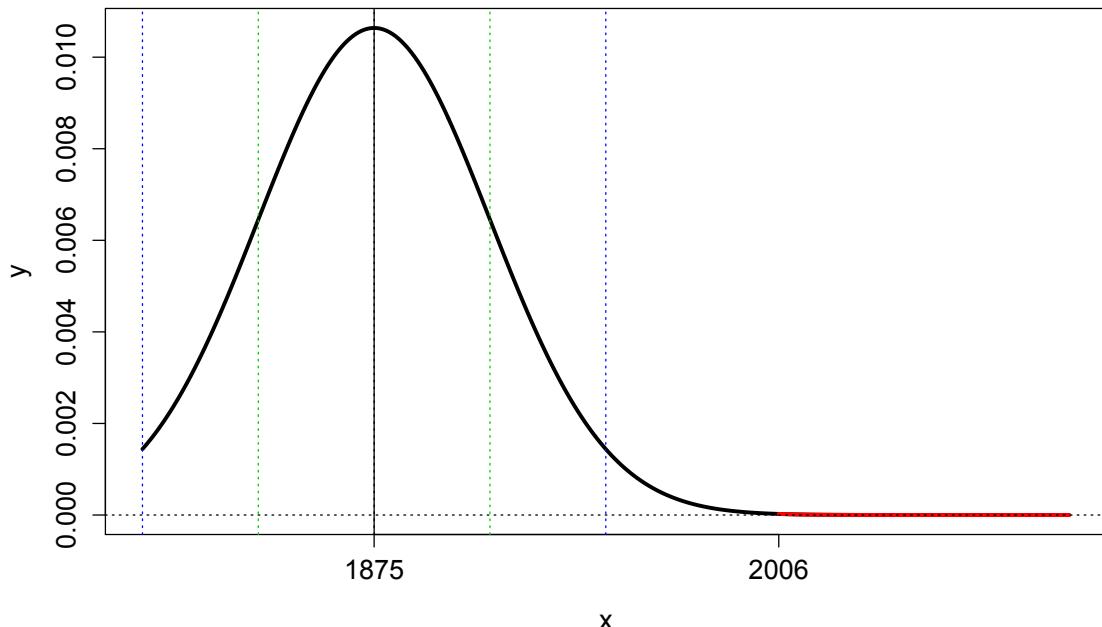


Figure 6.7: Normal approximation to the distribution of correct guesses in the Aquarius experiment, under the binomial hypothesis. The solid line is the mean, and the dotted lines show 1 and 2 SDs away from the mean.

## Lecture 7

# Confidence intervals and Normal Approximation

### 7.1 Confidence interval for sampling from a normally distributed population

In lecture 8 we learned about significance testing, which is one way to quantify uncertainty. Consider the following example: Data set #231 from [HDL<sup>+</sup>94] includes height measurements from 198 men and women, sampled at random from the much larger data 1980 OPCS (Office of Population Census and Survey) survey. The sample of 198 men's heights averages 1732mm, with an SD of 68.8mm. What does this tell us about the average height of British men?

We only measured 198 of the many millions of men in the country, but if we can assume that this was a random sample, this allows us to draw conclusions about the entire population from which it was sampled. We reason as follows: Imagine a box with millions of cards in it, on each of which is written the height of one British man. These numbers are normally distributed with mean  $\mu$  and variance  $\sigma^2$ . We get to look at 198 of these cards, sampled at random. Call the cards we see  $X_1, \dots, X_{198}$ . Our best guess for  $\mu$  will of course be the **sample mean**

$$\bar{X} = \frac{1}{198}(X_1 + \dots + X_{198}) = 1732\text{mm.}$$

“Best guess” is not a very precise statement, though. To make a precise statement, we use the **sampling distribution** of the estimation error  $\bar{X} - \mu$ . The average of independent normal random variables is normal. The

expectation is the average of the expectations, while the variance is

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var}(X_1 + \dots + X_{198}) = \frac{1}{n^2} (\text{Var}(X_1) + \dots + \text{Var}(X_{198})) = \frac{\sigma^2}{n}.$$

since the variance of a sum of independent variables is always the sum of the variances. Thus, we can standardise  $\bar{X}$  by writing

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

which is a standard normal random variable (that is, with expectation 0 and variance 1).

A tiny bit of algebra gives us

$$\mu = \bar{X} - \frac{\sigma}{\sqrt{n}} Z.$$

This expresses the unknown quantity  $\mu$  in terms of known quantities and a random variable  $Z$  with known distribution. Thus we may use our standard normal tables to generate statements like “the probability is 0.95 that  $Z$  is in the range  $-1.96$  to  $1.96$ ,” implying that

the probability is 0.95 that  $\mu$  is in the range  $\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}$  to  $\bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$ .

(Note that we have used the fact that the normal distribution is symmetric about 0.) We call this interval a **95% confidence interval** for the unknown population mean.

The quantity  $\sigma/\sqrt{n}$ , which determines the scale of the confidence interval, is called the **Standard Error** for the sample mean, commonly abbreviated SE. If we take  $\sigma$  to be the sample standard deviation — more about this assumption in chapter 10 — the Standard Error is  $69\text{mm}/\sqrt{198} \approx 4.9\text{mm}$ . The 95% confidence interval for the population mean is then  $1732 \pm 9.8\text{mm}$ , so  $(1722, 1742)\text{mm}$ . In place of our vague statement about a best guess for  $\mu$ , we have an interval of width 20 mm in which we are 95% confident that the true population mean lies.

**General procedure for normal confidence intervals:** Suppose  $X_1, \dots, X_n$  are independent samples from a normal distribution with unknown mean  $\mu$ , and known variance  $\sigma^2$ . Then a (symmetric)  $c\%$  normal confidence interval for  $\mu$  is the interval

$$(\bar{X} - z \text{SE}, \bar{X} + z \text{SE}), \text{ which we also write as } \bar{X} \pm z \text{SE},$$

where  $SE = \sigma/\sqrt{n}$ , and  $z$  is the appropriate quantile of the standard normal distribution. That is, it is the number such that  $(100 - c)/2\%$  of the probability in the standard normal distribution is above  $z$ . Thus, if we're looking for a 95% confidence interval, we take  $\bar{X} \pm 2SE$ , whereas a 99% confidence interval would be  $\bar{X} \pm 2.6SE$ , since we see on the normal table that  $P(Z < 2.6) = 0.9953$ , so  $P(Z > 2.6) = 0.0047 \approx 0.5\%$ . (Note: The central importance of the 95% confidence interval derives primarily from its convenient correspondence to a  $z$  value of 2. More precisely, it is 1.96, but we rarely need — or indeed, can justify — such such precision.)

Level	68%	90%	95%	99%	99.7%
$z$	1.0	1.64	1.96	2.6	3.0
Prob. above $z$	0.16	0.05	0.025	0.005	0.0015

Table 7.1: Parameters for some commonly used confidence intervals.

In other situations, as we will see, we use the same formula for a normal confidence interval for a parameter  $\mu$ . The only thing that changes from problem to problem is the point estimate  $\bar{X}$ , and the standard error.

## 7.2 Interpreting the confidence interval

But what does confidence mean? The quantity  $\mu$  is a fact, not a random quantity, so we cannot say “The probability is 0.95 that  $\mu$  is between 1722mm and 1742mm.”<sup>1</sup> The randomness is in our estimate  $\hat{\mu} = \bar{X}$ . The true probability statement  $\mathbb{P}\{\bar{X} \in (\mu - 1.96SE, \mu + 1.96SE)\} = 0.95$  is equivalent, by simple arithmetic, to  $\mathbb{P}\{\mu \in (\bar{X} - 1.96SE, \bar{X} + 1.96SE)\} = 0.95$ . The latter statement *looks like* something different, a probability statement about  $\mu$ , but really it is a probability statement about the random interval: 95% of the time, the random interval generated according to this recipe will cover the true parameter.

**Definition 7.2.1.** A  $(\alpha \times 100)\%$  confidence interval (also called a confidence interval with **confidence coefficient** or **confidence level**  $\alpha$ ) for a parameter  $\mu$ , based on observations  $\mathbf{X} := (X_1, \dots, X_n)$  is a pair of statistics

---

<sup>1</sup>An alternative approach to statistics, called *Bayesian statistics*, does allow us to make precise sense of probability statements about unknown parameters, but we will not be considering it in this course.

(that is, quantities you can compute from the data  $\mathbf{X}$ )  $A(\mathbf{X})$  and  $B(\mathbf{X})$ , such that

$$\mathbb{P}\{A(\mathbf{X}) \leq \mu \leq B(\mathbf{X})\} = \alpha.$$

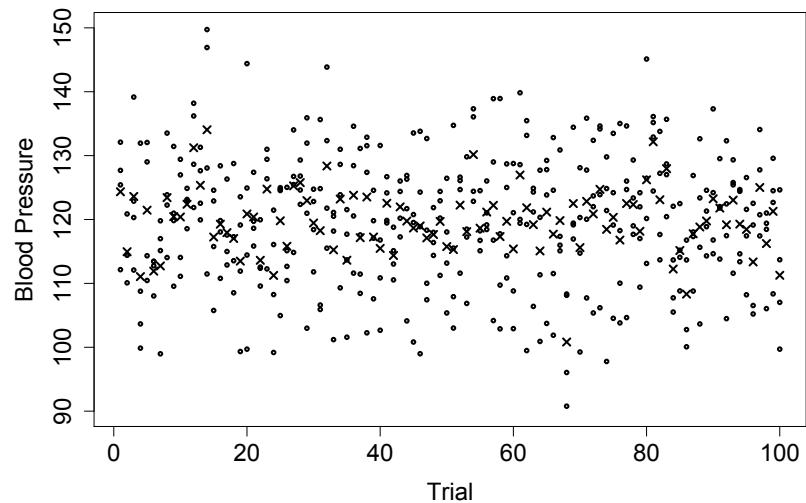
The quantity  $\mathbb{P}\{A(\mathbf{X}) \leq \theta \leq B(\mathbf{X})\}$  is called the **coverage probability** for  $\mu$ . Thus, a confidence interval for  $\mu$  with confidence coefficient  $\alpha$  is precisely a random interval with coverage probability  $\alpha$ . In many cases, it is not possible to find an interval with exactly the right coverage probability. We may have to content ourselves with an *approximate confidence interval* (with coverage probability  $\approx \alpha$ ) or a *conservative confidence interval* (with coverage probability  $\geq \alpha$ ). We usually make every effort not to overstate our confidence about statistical conclusions, which is why we try to err on the side of making the coverage probability — hence the interval — too large.

An illustration of this problem is given in Figure 7.1. Suppose we are measuring systolic blood pressure on 100 patients, where the true blood pressure is 120 mmHg, but the measuring device makes normally distributed errors with mean 0 and SD 10 mmHg. In order to reduce the errors, we take four measures on each patient and average them. Then we compute a confidence interval. The measures are shown in figure 7.1(a). In Figure 7.1(b) we have shown a 95% confidence interval for each patient, computed by taking the average of the patient's four measurements, plus and minus 10. Notice that there are 6 patients (shown by red X's for their means) where the true measure — 120 mmHg — lies outside the confidence interval. In Figures 7.1(c) and 7.1(d) we show 90% and 68% confidence intervals, which are narrower, and hence miss the true value more frequently.

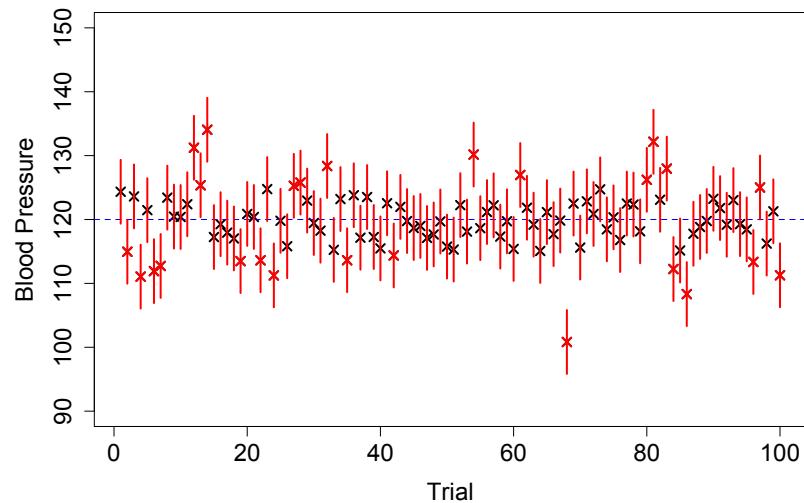
A 90% confidence interval tells you that 90% of the time the true value will lie in this range. In fact, we find that there are exactly 90 out of 100 cases where the true value is in the confidence interval. The 68% confidence intervals do a bit better than would be expected on average: 74 of the 100 trials had the true value in the 68% confidence interval.

Confidence intervals

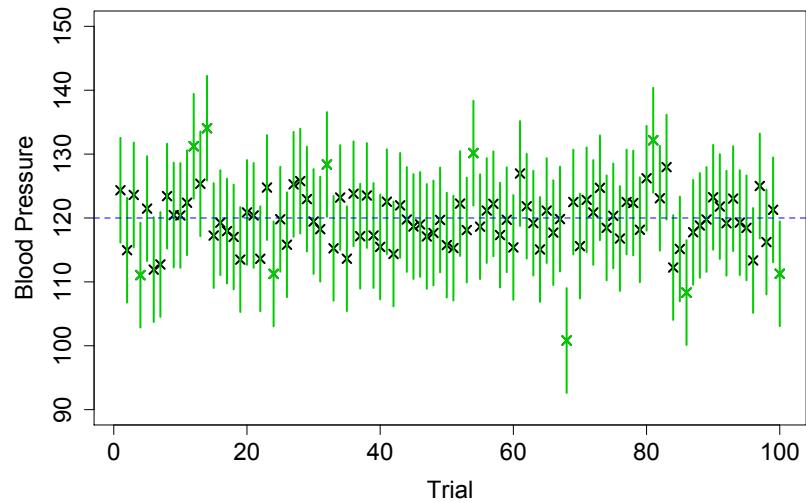
125



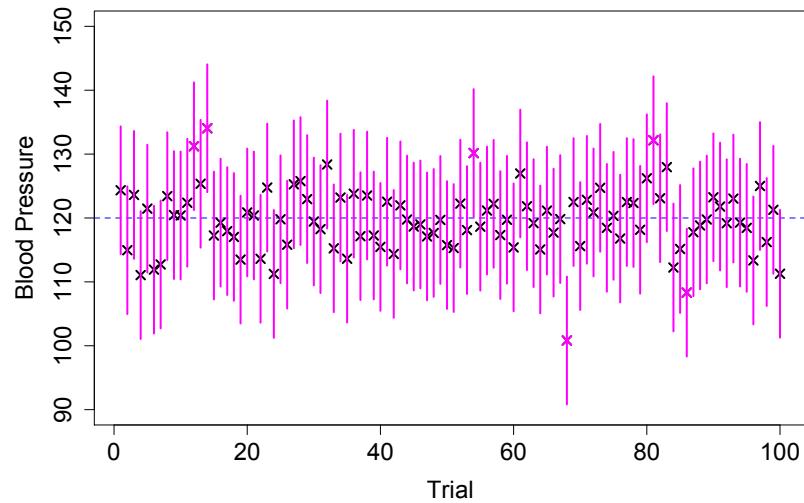
(a) 100 patients



(b) 95% confidence intervals



(c) 90% confidence intervals



(d) 68% confidence intervals

Figure 7.1: Confidence intervals for 100 patients' blood pressure, based on four measurements. Each column of Figure 7.1(a) shows a single patient's four measurements. The true BP in each case is 120, and the measurement errors are normally distributed with mean 0 and SD 10.

### 7.3 Confidence intervals for probability of success

We discussed in section 6.8 that the binomial distribution can be well approximated by a normal distribution. This means that if we are estimating the probability of success  $p$  from some observations of successes and failures, we can use the same methods as above to put a confidence interval on  $p$ . For instance, the Gallup organisation carried out a poll in October, 2005, of Americans' attitudes about guns (see <http://www.gallup.com/poll/20098/gun-ownership-use-america.aspx>). They surveyed 1,012 Americans, chosen at random. Of these, they found that 30% said they personally owned a gun. But, of course, if they'd picked different people, purely by chance they would have gotten a somewhat different percentage. How different could it have been? What does this survey tell us about the true fraction (call it  $p$ ) of Americans who own guns?

We can compute a 95% confidence interval as  $0.30 \pm 1.96 \text{ SE}$ . All we need to know is the SE for the proportion  $p$ , which is the same as the standard deviation for the *observed proportion of successes*. We know from section 6.8 (and discussed again at length in section 8.3), that the standard error is

$$\text{SE} = \sqrt{\frac{p(1-p)}{n}},$$

where  $n$  is the number of samples. In this case, we get  $\text{SE} = \sqrt{0.3 \times 0.7 / 1012} = 0.0144$ . So

95% confidence interval for  $p$  is  $0.30 \pm 0.029 = (0.271, 0.329)$ .

Loosely put, we can be 95% confident that the true proportion supporting EPP is between 27% and 33%. A 99% confidence interval comes from multiplying by 2.6 instead of 1.96: it goes from 26.3% to 33.7%.

Notice that the Standard Error for a proportion is a maximum when  $p = 0.5$ . Thus, we can always get a “conservative confidence interval” — an interval where the probability of finding the true parameter in it is **at least** 95% (or whatever the level is) by taking the SE to be  $\sqrt{.25/n}$ . The 95% confidence interval then has the particularly simple form sample mean  $\pm 1/\sqrt{n}$ .

## 7.4 The Normal Approximation

### Approximation Theorems in Probability

Suppose  $X_1, X_2, \dots, X_n$  are independent samples from a probability distribution with mean  $\mu$  and variance  $\sigma^2$ . Then

**Law of Large Numbers (LLN):** For  $n$  large,  $\bar{X}$  will be close to  $\mu$ .

**Central Limit Theorem (CLT):** For  $n$  large, the error in the LLN is close to a normal distribution, with variance  $\sigma^2/n$ . That is, using our standardisation procedure for the normal distribution,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (7.1)$$

is close to having a standard normal distribution. Equivalently,  $X_1 + \dots + X_n$  has approximately  $N(n\mu, n\sigma^2)$  distribution.

So far, we have been assuming that our data are sampled from a population with a normal distribution. What justification do we have for this assumption? And what do we do if the data come from a different distribution? One of the great early discoveries of probability theory is that many different kinds of random variables come close to a normal distribution when you average enough of them. You have already seen examples of this phenomenon in the normal approximation to the binomial distribution and the Poisson.

In probability textbooks such as [Fel71] you can find very precise statements about what it means for the distribution to be “close”. For our purposes, we will simply treat  $Z$  as being actually a standard normal random variable. However, we also need to know what it means for  $n$  to be “large”. For most distributions that you might encounter, 20 is usually plenty, while 2 or 3 are not enough. The key rules of thumb are that the approximation works best when the distribution of  $X$

- (i). is reasonably symmetric: Not skewed in either direction.

- (ii). has thin tails: Most of the probability is close to the mean, not many SDs away from the mean.

More specific indications will be given in the following examples.

### 7.4.1 Normal distribution

Suppose  $X_i$  are drawn from a normal distribution with mean  $\mu$  and variance  $\sigma^2 = N(\mu, \sigma^2)$ . We know that  $X_1 + \dots + X_n$  has  $N(n\mu, n\sigma^2)$  distribution, and so that  $\bar{X}$  has  $N(\mu, \sigma^2/n)$  distribution. A consequence is that  $Z$ , as defined in (7.1), in fact has exactly the standard normal distribution. In fact, this is an explanation for why the CLT works: The normal distribution is the only distribution such that when you average multiple copies of it, you get another distribution of the same sort.<sup>2</sup> Other distributions are not stable under averaging, and naturally converge to the distributions that are.

### 7.4.2 Poisson distribution

Suppose  $X_i$  are drawn from a Poisson distribution with parameter  $\mu$ . The variance is then also  $\mu$ . We know that  $X_1 + \dots + X_n$  has Poisson distribution with parameter  $n\mu$ . The CLT tells us, then, that for  $n$  large enough, the  $Po(n\mu)$  distribution is very close to the  $N(n\mu, n\mu)$  distribution; or, in other words,  $Po(\lambda)$  is approximately the same as  $N(\lambda, \lambda)$  for  $\lambda$  large. How large should it be?

The Poisson distribution is shown in Figure 7.2 for different values of  $\lambda$ , together with the approximating normal density curve. One way of seeing the failure of the approximation for small  $\lambda$  is to note that when  $\lambda$  is not much bigger than  $\sqrt{\lambda}$  — much bigger meaning a factor of 2.5 or so, so  $\lambda < 6.2$ , the normal curve will have substantial probability below  $-0.5$ . Since this is supposed to approximate the probability of the corresponding Poisson distribution below 0, this manifestly represents a failure. For instance, when  $\lambda = 1$ , the  $Po(1)$  distribution is supposed to be approximated by  $N(1, 1)$ , implying

$$P\{Po(1) < 0\} \approx P\{N(1, 1) < -0.5\} \approx P\{N(0, 1) < -1.5\} = .067.$$

In general, the threshold  $-0.5$  corresponds to  $Z = (-0.5 - \lambda)/\sqrt{\lambda}$ . The corresponding values for other parameters are given in Table 7.2.

---

<sup>2</sup>Technically, it is the only distribution *with finite variance* for which this is true.

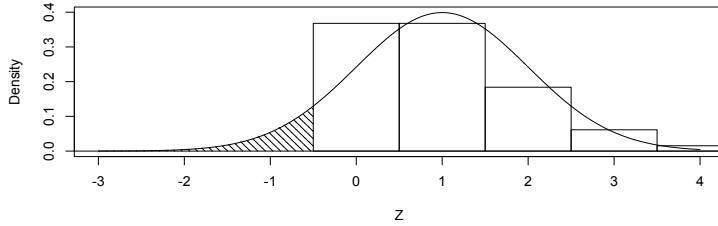
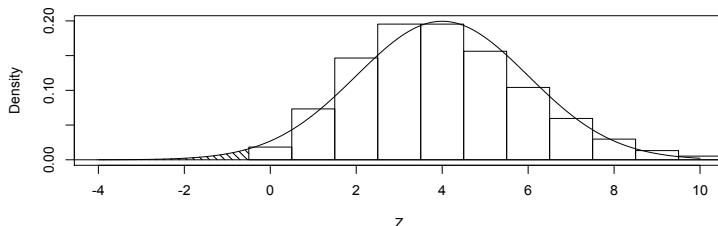
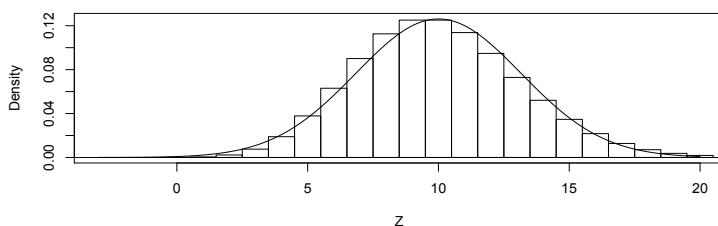
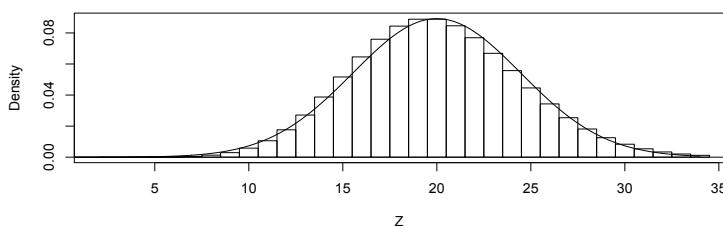
(a)  $\lambda = 1$ (b)  $\lambda = 4$ (c)  $\lambda = 10$ (d)  $\lambda = 20$ 

Figure 7.2: Normal approximations to  $Po(\lambda)$ . Shaded region is the implied approximate probability of the Poisson variable  $< 0$ .

$\lambda$	Standardised Z	Normal probability
1	-1.5	0.067
4	-2.25	0.012
10	-3.32	0.00045
20	-4.58	0.0000023

Table 7.2: Probability below  $-0.5$  in the normal approximation to Poisson random variables with different parameters  $\lambda$ .

### 7.4.3 Bernoulli variables

“Bernoulli variables” is the name for random variables that are 1 or 0, with probability  $p$  or  $1 - p$  respectively. Then  $B = X_1 + \dots + X_n$  is the number of successes in  $n$  trials, with success probability  $p$  each time — that is, a  $\text{Bin}(n, p)$  random variable. Again, we have already discussed that binomial random variables may be approximated by normal random variables.  $X_i$  has expectation  $p$  and variance  $p(1 - p)$ . The CLT then implies that  $\text{Bin}(n, p) \sim N(np, np(1 - p))$  for large values of  $n$ . Note that  $B/n$  is the proportion of successes in  $n$  trials, and this has approximately  $N(p, p(1 - p)/n)$  distribution. In other words, the observed proportion will be close to  $p$ , but will be off by a small multiple of the SD, which shrinks as  $\sigma/\sqrt{n}$ , where  $\sigma = \sqrt{p(1 - p)}$ . This is exactly the same thing we discussed in section 7.3.

How large does  $n$  need to be? As in the case of the Poisson distribution, discussed in section 7.4.2, a minimum requirement is that the mean be substantially larger than the SD; in other words,  $np \gg \sqrt{np(1 - p)}$ , so that  $n \gg 1/p$ . (The condition is symmetric, so we also need  $n \gg 1/(1 - p)$ .) This fits with our rule of thumb that  $n$  needs to be bigger when the distribution of  $X$  is skewed, which is the case when  $p$  is close to 0 or 1.

In Figure 7.3 we see that when  $p = 0.5$  the normal approximation is quite good, even when  $n$  is only 10; on the other hand, when  $p = 0.1$  we have a good normal approximation when  $n = 100$ , but not when  $n$  is 25. (Note, by the way, that  $\text{Binom}(25, 0.1)$  is approximately  $\text{Po}(2.5)$ , so this is closely related to the observations we made in section 7.4.2.)

## 7.5 CLT for real data

We show how the CLT is applied to understand the mean of samples from real data. It permits us to apply our Z and t tests for testing population

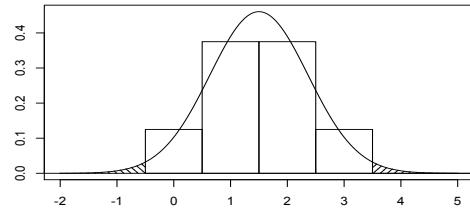
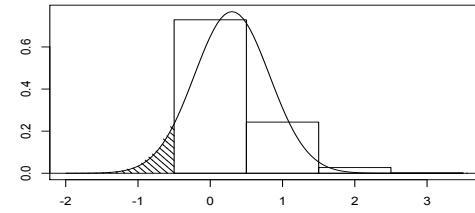
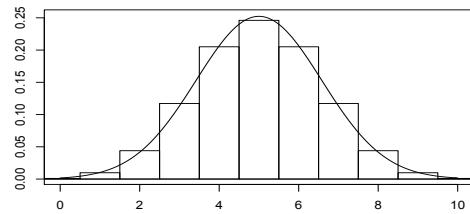
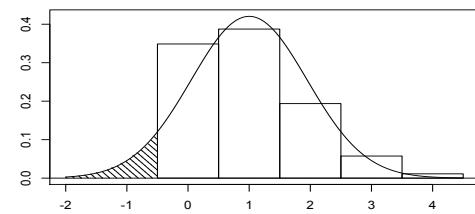
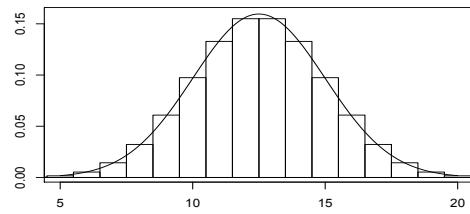
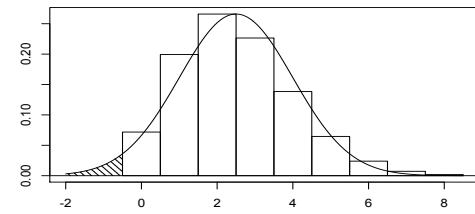
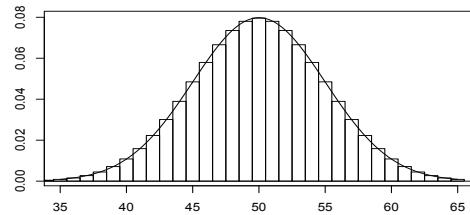
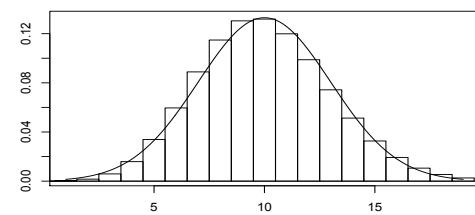
(a)  $p = 0.5, n = 3$ (b)  $p = 0.1, n = 3$ (c)  $p = 0.5, n = 10$ (d)  $p = 0.1, n = 10$ (e)  $p = 0.5, n = 25$ (f)  $p = 0.1, n = 25$ (g)  $p = 0.5, n = 100$ (h)  $p = 0.1, n = 100$ 

Figure 7.3: Normal approximations to  $\text{Binom}(n, p)$ . Shaded region is the implied approximate probability of the Binomial variable  $< 0$  or  $> n$ .

means and computing confidence intervals for the population mean (as well as for differences in means) even to data that are not normally distributed. (Caution: Remember that  $t$  is an improvement over  $Z$  only when the number of samples being averaged is small. Unfortunately, the CLT itself may not apply in such a case.) We have already applied this idea when we did the  $Z$  test for proportions, and the CLT was also hidden in our use of the  $\chi^2$  test.

### 7.5.1 Quebec births

We begin with an example that is well suited to fast convergence. We have a list of 5,113 numbers, giving the number of births recorded each day in the Canadian province of Quebec over a period of 14 years, from 1 January, 1977 through 31 December, 1990. (The data are available at the Time Series Data Library <http://www.robjhyndman.com/TSDL/>, under the rubric “demography”.) A histogram of the data is shown in Figure 7.4(a). The mean number of births is  $\mu = 251$ , and the SD is  $\sigma = 41.9$ .

Suppose we were interested in the average number of daily births, but couldn’t observe data for all of the days. How many days would we need to observe to get a reasonable estimate? Obviously, if we observed just a single day’s data, we would be seeing a random pick from the histogram 7.4(a), which could be far off of the true value. (Typically, it would be off by about the SD, which is 41.9.) Suppose we sample the data from  $n$  days, obtaining counts  $x_1, \dots, x_n$ , which average to  $\bar{x}$ . How far off might this be? The normal approximation tells us that a 95% confidence interval for  $\mu$  will be  $\bar{x} \pm 1.96 \cdot 41.9/\sqrt{n}$ . For instance, if  $n = 10$ , and we find the mean of our 10 samples to be 245, then a 95% confidence interval will be (229, 271). If there had been 100 samples, the confidence interval would be (237, 253). Put differently, the average of 10 samples will lie within 26 of the true mean 95% of the time, while the average of 100 samples will lie within 8 of the true mean 95% of the time.

This computation depends upon  $n$  being large enough to apply the CLT. Is it? One way of checking is to perform a simulation: We let a computer pick 1000 random samples of size  $n$ , compute the means, and then look at the distribution of those 1000 means. The CLT predicts that they should have a certain normal distribution, so we can compare them and see. If  $n = 1$ , the result will look exactly like Figure 7.4(a), where the curve in red is the appropriate normal approximation predicted by the CLT. Of course, there is no reason why the distribution should be normal for  $n = 1$ . We see that for  $n = 2$  the true distribution is still quite far from normal, but by  $n = 10$  the normal is already starting to fit fairly closely, and by  $n = 100$

the fit has become extremely good.

Suppose we sample 100 days at random. What is the probability that the total number of births is at least 25500? Let  $S = \sum_{i=1}^{100} X_i$ . Then  $S$  is normally distributed with mean 25100, and SD  $41.9\sqrt{100} = 419$ . We compute by standardising:

$$P\{S > 25500\} = P\left\{\frac{S - 25100}{419} > \frac{25500 - 25100}{419}\right\} = P\{Z > 0.95\}$$

where  $Z = (S - 25100)/419$ . By the CLT,  $Z$  has approximately the standard normal distribution, so we can look up its probabilities on the table, and see that  $P\{Z > 0.95\} = 1 - P\{Z \leq 0.95\} = 0.171$ .

**Bonus question:**  $S$  comes in whole number values, so shouldn't we have made the cutoff 25500.5? Or should it be 25499.5? If we want to answer the question about the probability that  $S$  is **strictly** bigger than 25500, then the cutoff should be 25500.5. If we want the probability that  $S$  is **strictly** bigger than 25500, then the cutoff should be 25499.5. If we don't have a specific preference, then 25500 is a reasonable compromise. Of course, in this case, it only makes a difference of about 0.002 in the value of  $Z$ , which is negligible.

This is of course the same as the probability that the average number of births is at least 255. We could also compute this by reasoning that  $\bar{X} = S/100$  is normally distributed with mean 251 and SD  $41.9/\sqrt{100} = 4.19$ . Thus,

$$P\{\bar{X} > 255\} = P\left\{\frac{\bar{X} - 251}{4.19} > \frac{255 - 251}{4.19}\right\} = P\{Z > 0.95\},$$

which comes out to the same thing.

### 7.5.2 California incomes

A standard example of a highly skewed distribution — hence a poor candidate for applying the CLT — is household income. The mean is much greater than the median, since there are a small number of extremely high incomes. It is intuitively clear that the average of incomes must be hard to predict. Suppose you were sampling 10,000 Americans at random — a very large sample — whose average income is £30,000. If your sample happens

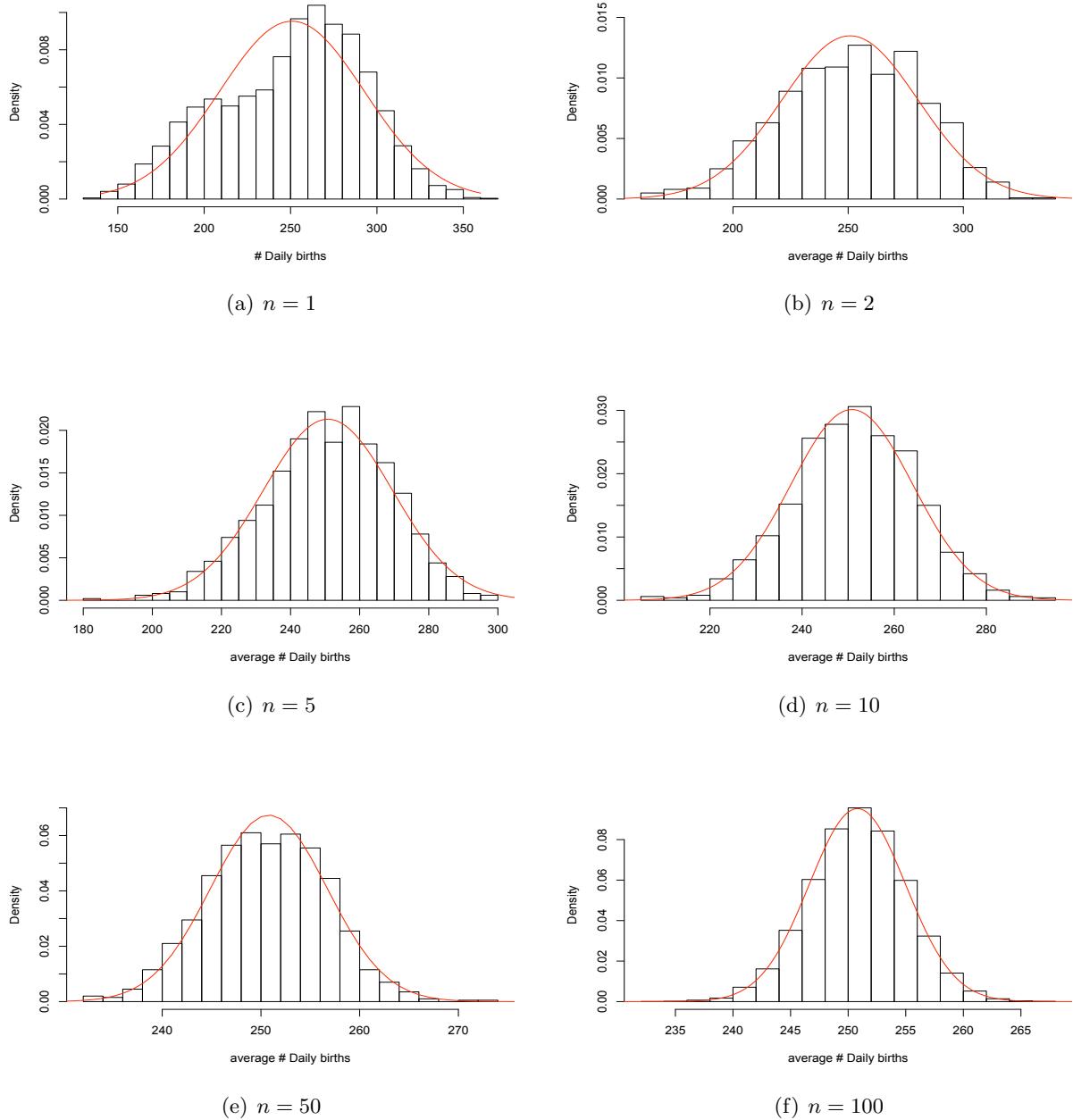


Figure 7.4: Normal approximations to averages of  $n$  samples from the Quebec birth data.

to include Bill Gates, with annual income of, let us say, £3 billion, then his income will be ten times as large as the total income of the entire remainder of the sample. Even if everyone else has zero income, the sample mean will be at least £300,000. The distribution of the mean will not converge, or will converge only very slowly, if it can be substantially affected by the presence or absence of a few very high-earning individuals in the sample.

Figure 7.5(a) is a histogram of household incomes, in thousands of US dollars, in the state of California in 1999, based on the 2000 US census (see [www.census.gov](http://www.census.gov)). We have simplified somewhat, since the final category is “more than \$200,000”, which we have treated as being the range \$200,000 to \$300,000. (Remember that histograms are on a density scale, with the area of a box corresponding to the number of individuals in that range. Thus, the last three boxes all correspond to about 3.5% of the population, despite their different heights.) The mean income is about  $\mu = \$62,000$ , while the median is \$48,000. The SD of the incomes is  $\sigma = \$55,000$ .

Figures 7.5(b)–7.5(f) show the effect of averaging 2,5,10,50, and 100 randomly chosen incomes, together with a normal distribution (in green) as predicted by the CLT, with mean  $\mu$  and variance  $\sigma^2/n$ . We see that the convergence takes a little longer than it did with the more balanced birth data of Figure 7.4 — averaging just 10 incomes is still quite skewed — but by the time we have reached the average of 100 incomes the match to the predicted normal distribution is remarkably good.

## 7.6 Using the Normal approximation for statistical inference

There are many implications of the Central Limit Theorem. We can use it to estimate the probability of obtaining a total of at least 400 in 100 rolls of a fair six-sided die, for instance, or the probability of a subject in an ESP experiment, guessing one of four patterns, obtaining 30 correct guesses out of 100 purely by chance. These were discussed in lecture 6 of the first set of lectures. It suggests an explanation for why height and weight, and any other quantity that is affected by many small random factors, should end up being normally distributed.

Here we discuss one crucial application: The CLT allows us to compute normal confidence intervals and apply the Z test to data that are not themselves normally distributed.

### 7.6.1 An example: Average incomes

Suppose we take a random sample of 400 households in Oxford, and find that they have an average income of £36,200, with an SD of £26,400. What can we infer about the average income of all households in Oxford?

**Answer:** Although the distribution of incomes is not normal — and if we weren't sure of that, we could see from the fact that the SD is not much smaller than the mean — the average of 400 incomes will be normally distributed. The SE for the mean is  $\text{£}26400/\sqrt{400} = 1320$ , so a 95% confidence interval for the average income in the population will be  $\text{£}36200 \pm 1.96 \cdot \text{£}1320 = (\text{£}33560, \text{£}38840)$ . A 99% confidence interval is  $\text{£}36200 \pm 2.6 \cdot \text{£}1320 = (\text{£}32800, \text{£}39600)$ .

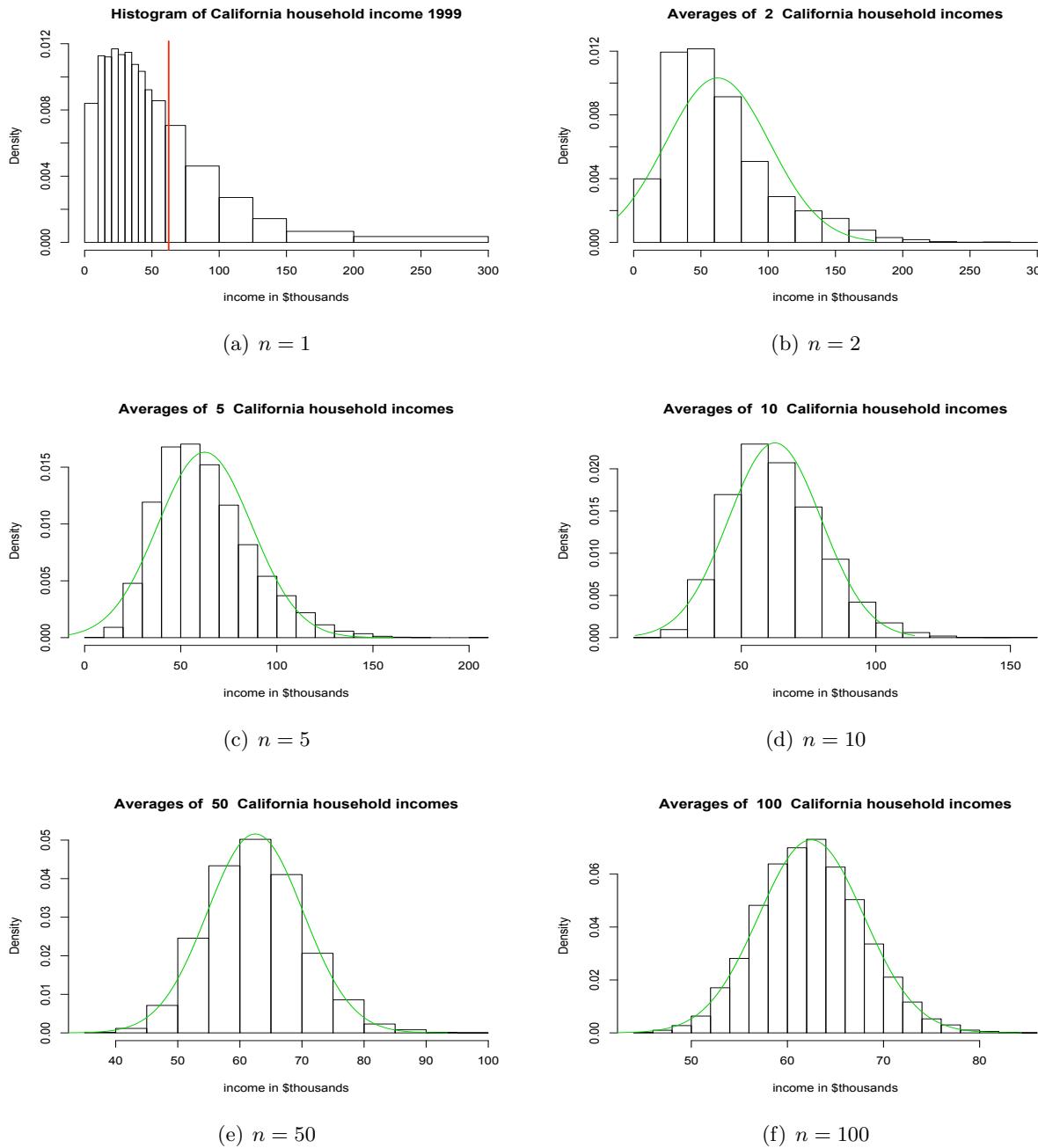


Figure 7.5: Normal approximations to averages of  $n$  samples from the California income data. The green curve shows a normal density with mean  $\mu$  and variance  $\sigma^2/n$ .



# Lecture 8

## The Z Test

### 8.1 Introduction

In Lecture 1 we saw that statistics has a crucial role in the scientific process and that we need a good understanding of statistics in order to avoid reaching invalid conclusions concerning the experiments that we do. In Lectures 2 and 3 we saw how the use of statistics necessitates an understanding of probability. This lead us to study how to calculate and manipulate probabilities using a variety of probability rules. In Lectures 4, 5 and 6 we consider three specific probability distributions that turn out to be very useful in practical situations. Effectively, all of these previous lectures have provided us with the basic tools we need to use statistics in practical situations.

The goal of statistical analysis is to draw reasonable conclusions from the data and, perhaps even more important, to give precise statements about the level of certainty that ought to be attached to those conclusions. In lecture 7 we used the normal distribution to derive one form that these “precise statements” can take: a confidence interval for some population mean. In this lecture we consider an alternative approach to describing very much the same information: Significance tests.

### 8.2 The logic of significance tests

#### Example 8.1: Baby-boom hypothesis test

Consider the following hypothetical situation: Suppose we think that UK newborns are heavier than Australian newborns. We

know from large-scale studies that UK newborns average 3426g, with an SD of 538g. (See, for example, [NNGT02].) The weights are approximately normally distributed. We think that maybe babies in Australia have a mean birth weight smaller than 3426g and we would like to test this hypothesis.

Intuitively we know how to go about testing our hypothesis. We need to take a sample of babies from Australia, measure their birth weights and see if the sample mean is *significantly smaller* than 3426g. Now, we have a sample of 44 Australian newborns, presented in Table 1.2, and with histogram presented in Figure 8.1. (Ignore for the moment that these are not really a sample of all Australian babies...)

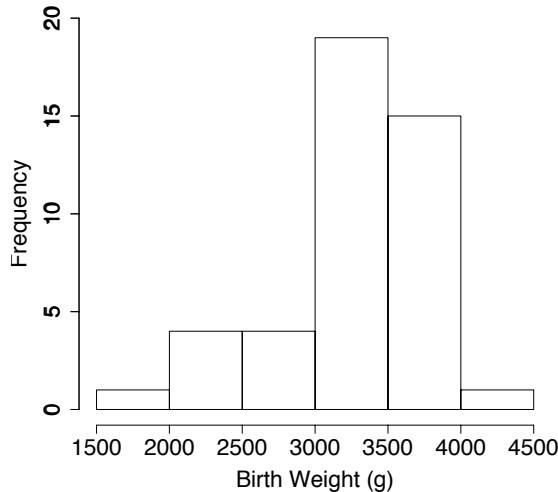


Figure 8.1: A Histogram showing the birth weight distribution in the Baby-boom dataset.

We observe that the sample mean of these 44 weights is 3276g. So we might just say that we're done. The average of these weights is smaller than 3426g, which is what we wanted to show.

“But wait!” a skeptic might say. “You might have just happened by chance to get a particularly heavy group of newborns. After all, even in England lots of newborns are lighter than 3276g. And 44 isn't such a big sample.”

How do we answer the skeptic? Is 44 big enough to conclude that there is a real difference in weights between the Australian sample and the known English average? We need to distinguish between

The research hypothesis: “Australian newborns have a mean weight greater than 3426g,”

and

The null hypothesis: “There’s no difference in mean weight; the apparent difference is purely due to chance.”

How do we decide which is true? We put the null hypothesis to the test. It says that the 44 observed weights are just like what you might observe if you picked 44 at random from the UK newborn population; that is, from a normal distribution with mean 3426g and SD 538g.

Let  $X_1, \dots, X_{44}$  be 44 weights picked at random from a  $\mathcal{N}(3426, 538^2)$  distribution, and let  $X = \frac{1}{44}(X_1 + \dots + X_{44})$  be their mean. How likely is it that  $X$  is as small as 3276? Of course, it’s never impossible, but we want to know how plausible it is.

We know from section 6.6 that

$$X_1 + \dots + X_{44} \sim \mathcal{N}(3426 \times 44, 538^2 \times 44), \text{ and}$$

$$X = \frac{1}{44}(X_1 + \dots + X_{44}) \sim \mathcal{N}(3426, 538^2/44) = \mathcal{N}(3000, 81^2).$$

Thus,

$$P(X \leq 3276) = P\left(Z \leq \frac{3276 - 3426}{81}\right) = P(Z \leq -1.81) = 1 - P(Z < 1.81),$$

where  $Z = (X - \mu)/\sigma = (X - 3426)/81$  has standard normal distribution. Looking this up on the standard normal table, we see that the probability is about 0.0351. ■

The probability 0.0351 that we compute at the end of Example 8.1 is called the **p-value** of the test. It tells us how likely it is that we would observe such an extreme result if the null hypothesis were true. The lower the p-value, the stronger the evidence *against* the null hypothesis. We are faced with the alternative: Either the null hypothesis is false, or we have by

chance happened to get a result that would only happen about one time in 30. This seems unlikely, but not impossible.

Pay attention to the double negative that we commonly use for significance tests: We have a research hypothesis, which we think would be interesting if it were true. We don't test it directly, but rather we use the data to challenge a less interesting **null hypothesis**, which says that the apparently interesting differences that we've observed in the data are simply the result of chance variation. We find out whether the data support the research hypothesis by showing that the null hypothesis is false (or unlikely). If the null hypothesis passes the test, then we know only that this particular challenge was inadequate. We haven't proven the null hypothesis. After all, we may just not have found the right challenger; a different experiment might show up the weaknesses of the null. (The potential strength of the challenge is called the "power" of the test, and we'll learn about that in section 13.2.)

What if the challenge succeeds? We can then conclude with confidence (how much confidence depends on the p-value) that the null was wrong. But in a sense, this is shadow boxing: We don't exactly know who the challenger is. We have to think carefully about what the plausible alternatives are. (See, for instance, Example 8.2.)

### 8.2.1 Outline of significance tests

The basic steps carried out in Example 8.1 are common to most significance tests:

- (i). Begin with a **research (alternative) hypothesis**.
- (ii). Set up the **null hypothesis**.
- (iii). Collect a sample of data.
- (iv). Calculate a **test statistic** from the sample of data.
- (v). Compare the test statistic to its **sampling distribution** under the null hypothesis and calculate the **p-value**. The strength of the evidence is larger, the smaller the p-value.

### 8.2.2 Significance tests or hypothesis tests? Breaking the .05 barrier

We use p-values to weigh scientific evidence. What if we need to make a decision?

One common situation is that the null hypothesis is being compared to an alternative that implies a definite course of action. For instance, we may be testing whether daily doses of vitamin C prevent colds: We take 100 subjects, and give them vitamin C supplements every day for a year, and no vitamin C supplement for another year, and compare the numbers of colds. At the end, we have two alternatives: either make a recommendation for vitamin C, or not.

The standard approach is to start by saying: The neutral decision is to make no recommendation, and we associate that with the null hypothesis, which says that any difference observed may be due to chance. In this system, the key goal is to control the likelihood of falsely making a positive recommendation (because we have rejected the null hypothesis). This situation, where we incorrectly reject the null hypothesis is called a **Type I Error**. The opposite situation, where we retain the null hypothesis although it is false, is called a **Type II Error**.

By definition, if the null hypothesis is true, the probability that the p-value is less than a given number  $\alpha$  is exactly  $\alpha$ . Thus, we begin our hypothesis test by fixing  $\alpha$ , the probability of a Type I error, to be some tolerably low number. We call this  $\alpha$  the **significance level** of the test. (A common choice is  $\alpha = 0.05$ , but the significance level can be anything you choose. If the consequences of a Type I Error would be extremely serious — for instance, if we are testing a new and very expensive cancer drug, with the expectation that we will move to prescribing this drug for all patients at great expense if it is shown to be “significantly” better — we might choose a smaller value of  $\alpha$ .)

In our current example, the p-value is about  $10^{-6}$  which is lower than 0.05. In this case, we would conclude that

“there is significant evidence against the null hypothesis at the 5% level”

Another way of saying this is that

“we reject the null hypothesis at the 5% level”

If the p-value for the test were much larger, say 0.23, then we would conclude that

Decision \ Truth	$H_0$ True	$H_0$ False
Retain $H_0$	Correct (Prob. $1 - \alpha$ )	Type II Error (Prob. = $\beta$ )
Reject $H_0$	Type I Error (Prob. = level = $\alpha$ )	Correct (Prob. = Power = $1 - \beta$ )

Table 8.1: Types of errors

“the evidence against the null hypothesis is not significant at the 5% level”

Another way of saying this is that

“we cannot reject the null hypothesis at the 5% level”

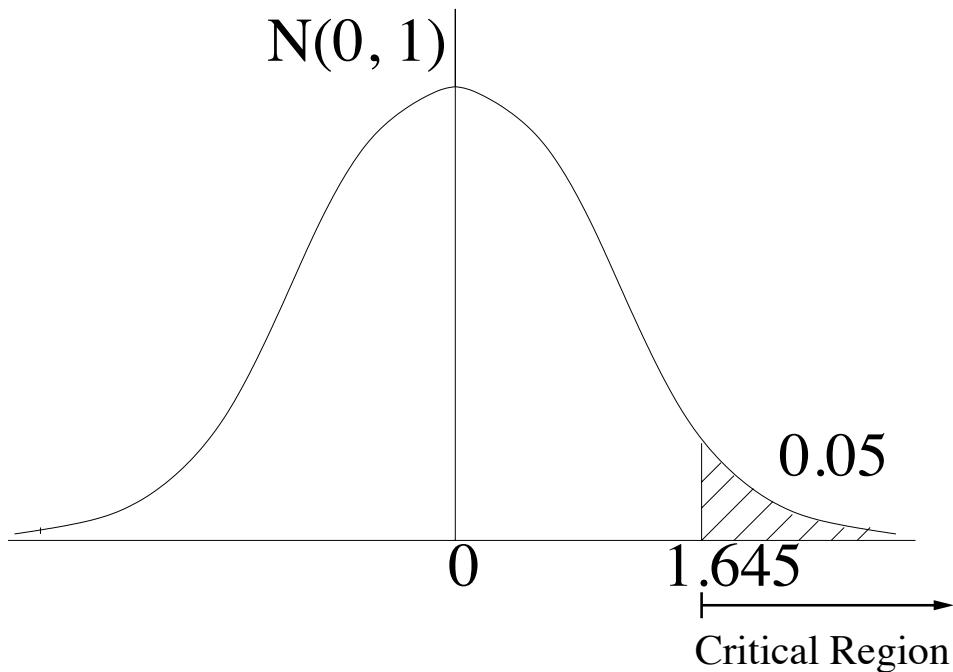
Note that the conclusion of a hypothesis test, strictly speaking, is binary: We either reject or retain the null hypothesis. There are no gradations, no strong rejection or borderline rejection or barely retained. The fact that our p-value was  $10^{-6}$  ought not to be taken, retrospectively, as stronger evidence against the null hypothesis than a p-value of 0.04 would have been.

By the strict logic imposed the data are completely used up in the test: If we are testing at the 0.05 level and the p-value is 0.06, we cannot then collect more data to see if we can get a lower p-value. We would have to throw away the data, and start a new experiment. Needless to say, this is not what scientists really do, which makes even the apparently clear-cut yes/no decision set-up of the hypothesis test in reality rather difficult to interpret.

It is also quite common to confuse this situation with using significance tests to judge scientific evidence. So common, in fact, that many scientific journals impose the 0.05 significance threshold to decide whether results are worth publishing. An experiment that resulted in a statistical test with a p-value of 0.10 is considered to have failed, even if it may very well be providing reasonable evidence of something important; if it resulted in a statistical test with a p-value of 0.05 then it is a success, even if the effect size is minuscule, and even though 1 out of 20 true null hypotheses will fail the test at significance level 0.05.

Another way of thinking about hypothesis tests is that there is some **critical region** of values such that if the test statistic lies in this region then we will reject  $H_0$ . If the test statistic lies outside this region we will not reject  $H_0$ . In our example, using a 5% level of significance this set of

values will be the most extreme 5% of values in the right hand tail of the distribution. Using our tables backwards we can calculate that the boundary of this region, called the **critical value**, will be 1.645. The value of our test statistic is 3.66 which lies in the critical region so we reject the null hypothesis at the 5% level.



### 8.2.3 Overview of Hypothesis Testing

Hypothesis tests are identical to significance tests, except for the choice of a *significance level* at the beginning, and the nature of the conclusions we draw at the end:

- (i). Begin with a **research (alternative) hypothesis** and decide upon a **level of significance** for the test.
- (ii). Set up the **null hypothesis**.
- (iii). Collect a sample of data.
- (iv). Calculate a **test statistic** from the sample of data.

- (v). Compare the test statistic to its **sampling distribution** under the null hypothesis and calculate the **p-value**,

*or equivalently,*

Calculate the **critical region** for the test.

- (vi). Reject the null hypothesis if

the p-value is less than the **level of significance**,

*or equivalently,*

the test statistic lies in the **critical region**.

Otherwise, retain the null hypothesis.

### 8.3 The one-sample Z test

A common situation in which we use hypothesis tests is when we have multiple independent observations from a distribution with unknown mean, and we can make a test statistic that is normally distributed. The null hypothesis should then tell us what the mean and standard error are, so that we can normalise the test statistic. The normalised test statistic is then commonly called  $Z$ . We always define  $Z$  by

$$Z = \frac{\text{observation} - \text{expectation}}{\text{standard error}}. \quad (8.2)$$

The expectation and standard error are the mean and the standard deviation of the *sampling distribution*: that is, the mean and standard deviation that the observation has when seen as a random variable, whose distribution is given by the null hypothesis. Thus,  $Z$  has been *standardised*: its distribution is standard normal, and the p-value comes from looking up the observed value of  $Z$  on the standard normal table.

We call this a “one-sample” test because we are interested in testing the mean of samples from a single distribution. This is as opposed to the “two-

sample test" (discussed in section ??), in which we are testing the difference in means between two populations.

### 8.3.1 Test for a population mean $\mu$

We know from Lecture 6 that if

$$X_1 \sim \mathcal{N}(\mu, \sigma^2) \quad X_2 \sim \mathcal{N}(\mu, \sigma^2)$$

then

$$\begin{aligned} \bar{X} = \frac{1}{2}X_1 + \frac{1}{2}X_2 &\sim \mathcal{N}\left(\frac{1}{2}\mu + \frac{1}{2}\mu, \left(\frac{1}{2}\right)^2\sigma^2 + \left(\frac{1}{2}\right)^2\sigma^2\right) \\ \Rightarrow \bar{X} &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2}\right) \end{aligned}$$

In general,

If  $X_1, X_2, \dots, X_n$  are  $n$  independent and identically distributed random variables from a  $\mathcal{N}(\mu, \sigma^2)$  distribution then

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Thus, if we are testing the null hypothesis

$$H_0 : \text{The } X_i \text{ have } \mathcal{N}(\mu, \sigma) \text{ distribution,}$$

the expectation is  $\mu$ , and the standard error is  $\sigma/\sqrt{n}$ . Thus,

When testing the sample mean of  $n$  normal samples, with known SD  $\sigma$ , for the null hypothesis mean=  $\mu$ , the test statistic is

$$Z = \frac{\text{sample mean} - \mu}{\sigma/\sqrt{n}}.$$

Thus, under the assumption of the null hypothesis the sample mean of 44 values from a  $\mathcal{N}(3426, 538^2)$  distribution is

$$\bar{X} \sim \mathcal{N}\left(3426, \frac{538^2}{44}\right) = \mathcal{N}(3426, 81^2)$$

### 8.3.2 Test for a sum

Under some circumstances it may seem more intuitive to work with the sum of observations rather than the mean. If  $S = X_1 + \dots + X_n$ , where the  $X_i$  are independent with  $\mathcal{N}(\mu, \sigma^2)$  distribution, then  $S \sim \mathcal{N}(n\mu, n\sigma^2)$ . That is, the expectation is  $n\mu$  and the standard error is  $\sigma\sqrt{n}$ .

When testing the sum of  $n$  normal samples, with known SD  $\sigma$ , for the null hypothesis  $\text{mean} = \mu$ , the test statistic is

$$Z = \frac{\text{observed sum of samples} - n\mu}{\sigma \times \sqrt{n}}.$$

### 8.3.3 Test for a total number of successes

Suppose we are observing independent trials, each of which has unknown probability of success  $p$ . We observe  $X$  successes. We have the estimate  $\hat{p} = X/n$ . Suppose we have some possible value  $p_0$  of interest, and we wish to test the null hypothesis

$$H_0 : p = p_0$$

against the alternative

$$H_1 : p > p_0.$$

We already observed in section 6.8 that the random variable  $X$  has distribution very close to normal, with mean  $pn$  and standard error  $\sqrt{np(1-p)}$ , as long as  $n$  is reasonably large. We have then the test statistic

When testing the number of successes in  $n$  trials, for the null hypothesis  $P(\text{success}) = p_0$ , the test statistic is

$$Z = \frac{\text{observed number of successes} - np_0}{\sqrt{np_0(1-p_0)}}.$$

#### Example 8.2: The Aquarius Machine, continued

We repeat the computation of Example 6.9. The null hypothesis, corresponding to “no extrasensory powers”, is

$$H_0 : p = p_0 = 0.25;$$

the alternative hypothesis, Tart's research hypothesis, is

$$H_1 : p > 0.25.$$

With  $n = 7500$ , the expected number of successes under the null hypothesis is  $7500 \times \frac{1}{4} = 1875$ , and the standard error is  $\sqrt{7500 \times \frac{1}{4} \times \frac{3}{4}} = 37.5$ . We compute the test statistic

$$\begin{aligned} Z &= \frac{\text{observed number of successes} - \text{expected number of successes}}{\text{standard error}} \\ &= \frac{2006 - 1875}{37.5} \\ &= 3.49. \end{aligned}$$

(This is slightly different from the earlier computation ( $z = 3.48$ ) because we conventionally ignore the continuity correction when computing test statistics.) Thus we obtain from the standard normal table a p-value of 0.0002.

So it is extraordinary unlikely that we would get a result this extreme purely by chance, if the null hypothesis holds. If  $p_0 = 1/4$ , then Tart happened to obtain a result that one would expect to see just one time in 5000. Must we then conclude that  $p_0 > 1/4$ ? And must we then allow that at least some subjects had precognitive powers? Actually, in this case we know what happened to produce this result. It seems that there were defects in the random number generator, making the same light less likely to come up twice in a row. Subjects presumably cued in to this pattern after a while — they were told after each guess whether they'd been right — and made use of it for their later guesses. Thus, the binomial distribution did not hold — the outcomes of different tries were not independent, and did not all have probability  $1/4$  — but not in the way that Tart supposed. Thus, one needs always to keep in mind: Statistical tests tell us that the come from our chance model, but it doesn't necessarily follow that our favourite alternative is true.

Some people use the term **Type III error** to refer to the mistake of correctly rejecting the null hypothesis, but for the wrong reason. Thus, to infer that the subjects had extrasensory powers from these data would have been a Type III error. ■

### 8.3.4 Test for a proportion

When testing for probability of success in independent trials, it often seems natural to consider the *proportion* of successes rather than the number of successes as the fundamental object. Under the null hypothesis

$$H_0 : p = p_0$$

the expected proportion of successes  $X/n$  is  $p_0$ , and the standard error is  $\sqrt{p_0(1 - p_0)/n}$ .

When testing the proportion of successes in  $n$  trials, for the null hypothesis  $P(\text{success}) = p_0$ , the test statistic is

$$Z = \frac{\text{proportion of successes} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

$Z$  has standard normal distribution.

The test statistic will come out exactly the same, regardless of whether we work with numbers of successes or proportions.

#### Example 8.3: The Aquarius machine, again

We repeat the computations of Example 8.2, treating the proportion of correct guesses as the basic object. The observed proportion of successes is  $\hat{p} = k/n = 2006/7500 = 0.26747$ . The standard error for the proportion is

$$SE_{\hat{p}} = \sqrt{p_0(1 - p_0)/n} = \sqrt{\frac{1}{4} \cdot \frac{3}{4}/7500} = 0.005.$$

Thus, the test statistic is

$$Z = \frac{0.26747 - 0.25}{0.005} = 3.49,$$

which is exactly the same as what we computed before. ■

#### Example 8.4: GBS and swine flu vaccine

In Example 5.7 we fit a Poisson distribution to the number of GBS cases by week after vaccination. We noted that the fit, given in Table 5.3, didn't look very good, and concluded that GBS cases were not independent of the time of vaccination. But we did not test this goodness of fit formally.

In the current formalism, the null hypothesis formalises the notion that GBS is independent of vaccination, so that numbers of GBS cases are Poisson distributed, with parameter  $\lambda = 1.33$ . We test this by looking at the number of weeks with 0 GBS cases, which was observed to be 16. The formal null hypothesis is

$$H_0 : P(0 \text{ cases in a week}) = e^{-1.33} = 0.2645.$$

The alternative hypothesis is

$$H_1 : P(0 \text{ cases in a week}) \neq 0.2645.$$

The observed proportion is  $16/40 = 0.4$ . The standard error is

$$SE = \sqrt{0.2645 \times 0.7355/40} = 0.0697.$$

Thus, we may compute

$$Z = \frac{0.2645 - 0.4}{0.0697} = -1.94$$

Looking this up on the table, we see that  $P(Z < -1.94) = 1 - P(Z < 1.94) = 0.026$ . Since we have a two-sided alternative, the p-value is twice this, or 0.052.

If we were doing a significance test at the 0.05 level (or any lower level), we would simply report that the result was not significant at the 0.05 level, and retain the null hypothesis. Otherwise, we simply report the p-value and let the reader make his or her own judgement. ■

### 8.3.5 General principles: The square-root law

The fundamental fact which makes statistics work is the fact that when we add up  $n$  independent observations, the expected value increases by a factor of  $n$ , while the standard error increases only by a factor of  $\sqrt{n}$ . Thus, when we divide by  $n$  to obtain a mean (or a proportion), the standard error ends up

shrinking by a factor of  $\sqrt{n}$ . This corresponds to our intuition that averaging many independent samples will tend to be closer to the true value than any single measurement. If the standard deviation of the population is  $\sigma$ , the standard error of the sample mean is  $\sigma/\sqrt{n}$ . Intuitively, the standard error tells us about how far off the sample mean will be from the true population mean (or true probability of success): we will almost never be off by more than 3 SEs.

## 8.4 One and two-tailed tests

In Example 8.1 we wanted to test the research hypothesis that mean birth weight of Australian babies was less than 3426g. This suggests that we had some prior information that the mean birth weight of Australian babies was definitely not higher than 3426g, and that the interesting question was whether the weight was lower. If this were not the case then our research hypothesis would be that the mean birth weight of Australian babies was different from 3426g. This allows for the possibility that the mean birth weight could be less than or greater than 3426g.

In this case we would write our hypotheses as

$$\begin{aligned} H_0 &: \mu = 3426\text{g} \\ H_1 &: \mu \neq 3426\text{g} \end{aligned}$$

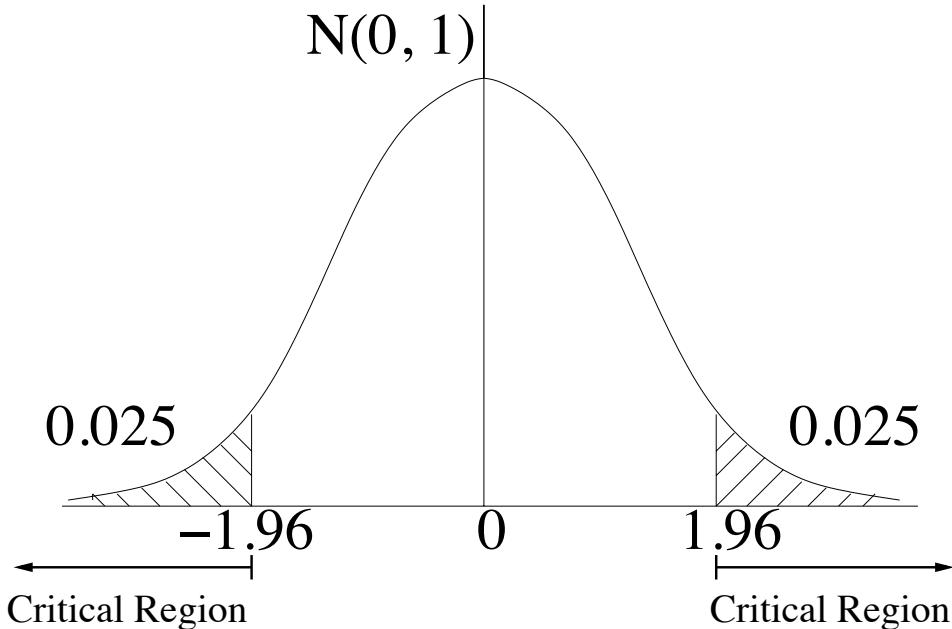
As before we would calculate our test statistic as  $-1.81$ . The p-value is different, though. We are not looking at the probability that  $Z$  is only less than  $-1.81$  (in the positive direction), but that  $Z$  is at least this big *in either direction*; so  $P(|Z| > 3.66)$ , or  $P(Z > 3.66) + P(Z < -3.66) = 2P(Z > 3.66)$ . Because of symmetry,

For a Z test, the two-tailed p-value is always twice as big as the one-tailed p-value.

In this case we allow for the possibility that the mean value is greater than 3426g by setting our critical region to be lowest 2.5% and highest 2.5% of the distribution. In this way the total area of the critical region remains 0.05 and so the level of significance of our test remains 5%. In this example, the critical values are  $-1.96$  and  $1.96$ . Thus if our test statistic is less than  $-1.96$  or greater than  $1.96$  we would reject the null hypothesis. In this example, the value of test statistic does lie in the critical region so we reject

the null hypothesis at the 5% level.

This is an example of a **two-sided test** as opposed to the previous example which was a **one-sided test**. The prior information we have in a specific situation dictates what we use as our alternative hypothesis which in turn dictates the type of test that we use.



Fundamentally, though, the distinction between one-tailed and two-tailed tests is important only because we set arbitrary p-values such as 0.05 as hard cutoffs. We should be cautious about “significant” results that depend for their significance on the choice of a one-tailed test, where a two-tailed test would have produced an “insignificant” result.

## 8.5 Hypothesis tests and confidence intervals

You may notice that we do a lot of the same things to carry out a statistical test that we do to compute a confidence interval. We compute a standard error and look up a value on a standard normal table. For instance, in Example 8.1 we might have expressed our uncertainty about the average Australian birthweight in the form of a confidence interval. The calculation would have been just slightly different:

We start with the mean observed birthweight in the Australian sample: 3276g. The standard error is  $\sigma/\sqrt{44}$ , where  $\sigma$  is the (unknown) SD of Australian birthweights. Since we don't know  $\sigma$ , we substitute the SD of the sample, which is  $s = 528g$ . So we use

$$SE = \sqrt{528g}\sqrt{44} = 80g.$$

Then a 95% confidence interval for the mean birthweight of Australian babies is  $3276 \pm 1.96 \cdot 80g = (3120, 3432)g$ ; a 99% confidence interval would be  $3276 \pm 2.58 \cdot 80g = (3071, 3481)g$ . (Again, remember that we are making the — not particularly realistic — assumption that the observed birthweights are a random sample of all Australian birthweights.) This is consistent with the observation that the Australian birthweights would just barely pass a test for having the same mean as the UK average 3426g, at the 0.05 significance level.

In fact, it is almost true to say that the symmetric 95% confidence interval contains exactly the possible means  $\mu_0$  such that the data would pass a test at the 0.05 significance level for having mean equal to  $\mu_0$ . What's the difference? It's in how we compute the standard error. In computing a confidence interval we estimate the parameters of the distribution from the data. When we perform a statistical test, we take the parameters (as far as possible) from the null hypothesis. In this case, that means that it makes sense to test based on the presumption that the standard deviation of weights is the SD of the UK births, which is the null hypothesis distribution. In this case, this makes only a tiny difference between 538g (the UK SD) and 528g (the SD of the Australian sample).

# Lecture 9

## The $\chi^2$ Test

### 9.1 Introduction — Test statistics that aren't Z

In lecture 8 we learned how to test hypotheses that a population has a certain mean, or that the probability of an event has a certain value. We base our test on the Z statistic, defined as

$$Z = \frac{\text{observed} - \text{expected}}{\text{standard error}},$$

which we then compare to a standard normal distribution. This  $Z$  has three properties that are crucial for making a test statistic:

- (i). We compute  $Z$  from the data;
- (ii). Extreme values of  $Z$  correspond to what we intuitively think of as important failures of the null hypothesis. The larger  $Z$  is, the more extremely the null hypothesis has failed;
- (iii). Under the null hypothesis, we know the distribution of  $Z$ .

In this lecture and many of the following ones we will learn about other statistical tests, for testing other sorts of scientific claims. The procedure will be largely the same: Formulate the claim in terms of the truth or falsity of a null hypothesis, find an appropriate test statistic (according to the principles enumerated above), and then judge the null hypothesis according to whether the p-value you compute is high (good for the null!) or low (bad for the null, good for the alternative!).

The Z test was used for testing whether the mean of quantitative data could have a certain value. In this lecture we consider categorical data.

These don't have a mean. We're usually interested in some claim about the distribution of the data among the categories. The most basic tool we have for testing whether data we observe really could have come from a certain distribution is called the  $\chi^2$  test.

### Example 9.1: Suicide and month of birth

A recent paper [SCB06] attempts to determine whether people born in certain months have higher risk of suicide than people born in other months. They gathered data for 26,915 suicides in England and Wales between 1979 and 2001, by people born between 1955 and 1966, for which birth dates were recorded. Looking just at the women in the sample, the data are summarised in Table 9.1, where the second column gives the fraction of dates that are in that month (that is, 31 or 30 or 28.25 divided by 365.25). The number of suicides is not the same every month

Month	Prob	Female	Male	Total
Jan	0.0849	527	1774	2301
Feb	0.0773	435	1639	2074
Mar	0.0849	454	1939	2393
Apr	0.0821	493	1777	2270
May	0.0849	535	1969	2504
Jun	0.0821	515	1739	2254
Jul	0.0849	490	1872	2362
Aug	0.0849	489	1833	2322
Sep	0.0821	476	1624	2100
Oct	0.0849	474	1661	2135
Nov	0.0821	442	1568	2010
Dec	0.0849	471	1690	2161
Total	1.0000	5801	21085	26886
Mean		483.4	1757.1	2240.5

Table 9.1: Suicides 1979–2001 by month of birth in England and Wales 1955–66, taken from [SCB06].

— as you would inevitably expect, by chance variation. But is there a pattern? Are there some months whose newborns are

more likely than others' to take their own lives 20 to 40 years later? There seem to be more suicides among spring babies. We might formulate a hypothesis as follows:

$$H_0 : \text{a suicide has equal chances } (1/365.25) \text{ of having been born on any date of the year}$$

The alternative hypothesis is that people born on some dates are more likely to commit suicide. (We count 29 February as 0.25 days, since it is present in one year out of four.)

How might we test this? One way would be to split up the suicides into two categories: "Spring births" (March through June) and "Others". We then have 9421 suicides among the spring births, and 17465 among the others. Now, 122 days are in the spring-birth category, so our null hypothesis is

$$H_0 : p_0 = \text{Probability of a suicide being spring-birth} = 122/365.25 = 0.334.$$

The expected number of spring-birth suicides (under the null hypothesis) is  $26886 \times p_0 = 8980$ , and the standard error is  $\sqrt{p_0(1 - p_0)n} = \sqrt{0.334 \times 0.666 \times 26886} = 77.3$ . We then perform a  $Z$  test with

$$Z = \frac{9421 - 8980}{77.3} = 5.71.$$

The  $Z$  statistic runs off the end of our table, indicating a p-value below 0.0001. (In fact, the p-value is below  $10^{-8}$ , or about 1 chance in 100 million.)

Is this right, though? Not really. We are guilty here of **data snooping**: We looked at the data in order to choose which way to break up the year into two categories. If we had split up the year differently, we might have obtained a very different result. (For example, if we had defined "spring-birth" to include only April through June, we would have obtained a  $Z$  statistic of only 0.415.) It would be helpful to have an alternative approach that could deal with multiple categories as they are, without needing to group them arbitrarily into two. ■

## 9.2 Goodness-of-Fit Tests

We are considering the following sort of situation. We observe  $n$  realisations of a categorical random variable. There are  $k$  categories. We have a null

hypothesis that tells us that these categories have probabilities  $p_1, \dots, p_k$ , so that the expected number of observations in the categories are  $np_1, \dots, np_k$ . The observed numbers in each category are  $n_1, \dots, n_k$ .

We want to have a test statistic that measures how far off the observed are, in total, from the expected. We won't fall into the trap of summing up  $n_i - np_i$  (which will always be 0). We might instead add up the squared differences  $(n_i - np_i)^2$ . But that seems like a problem, too: If we were expecting to have just 1 outcome in some category, and we actually got 11, that seems a lot more important than if we were expecting 1000 and actually got 1010, even though the difference is 10 in each case. So we want a difference to contribute more to the test statistic if the expected number is small.

By this reasoning we arrive at the  $\chi^2$  ("chi-squared", pronounced "keye squared") statistic

$$X^2 := \frac{(n_1 - np_1)^2}{np_1} + \dots + \frac{(n_k - np_k)^2}{np_k} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}. \quad (9.1)$$

### 9.2.1 The $\chi^2$ distribution

The statistic  $X^2$  has properties (i) and (ii) for a good test statistic: we can compute it from the data, and bigger values of  $X$  correspond to data that are farther away from what you would expect from the null hypothesis. But what about (iii)? We can't do a statistical test unless we know the distribution of  $X$  under the null hypothesis. Fortunately, the distribution is known. The statistic  $X$  has approximately (for large  $n$ ) one of a family of distributions, called the  $\chi^2$  distribution. There is a positive integer, called the **number of degrees of freedom** (abbreviated "d.f.") which tells us which  $\chi^2$  distribution we are talking about. One of the tricky things about using the  $\chi^2$  test statistic is figuring out the number of degrees of freedom. This depends on the number of categories, and how we picked the null hypothesis that we are testing. In general,

$\text{degrees of freedom} = \# \text{ categories} - \# \text{ parameters fit from the data} - 1.$

When is  $n$  large enough? The rule of thumb is that the expected number in every category must be at least about 5. So what do we do if some of the expected numbers are too small? Very simple: We group categories together,

until the problem disappears. We will see examples of this in sections 9.4.1 and 9.4.2.

The  $\chi^2$  distribution with  $d$  degrees of freedom is a continuous distribution<sup>1</sup> with mean  $d$  and variance  $2d$ . In Figure 9.1 we show the density of the chi-squared distribution for some choices of the degrees of freedom. We note that these distributions are always right-skewed, but the skew decreases as  $d$  increases. For large  $d$ , the  $\chi^2$  distribution becomes close to the normal distribution with mean  $d$  and variance  $2d$ .

As with the standard normal distribution, we rely on standard tables with precomputed values for the  $\chi^2$  distribution. We could simply have a separate table for each number of degrees of freedom, and use these exactly like the standard normal table for the  $Z$  test. This would take up quite a bit of space, though. (Potentially infinite — but for large numbers of degrees of freedom see section 9.2.2.) Alternatively, we could use a computer programme that computes p-values for arbitrary values of  $X$  and d.f. (In the R programming language the function `pchisq` does this.) This is an ideal solution, except that you don't have computers to use on your exams.

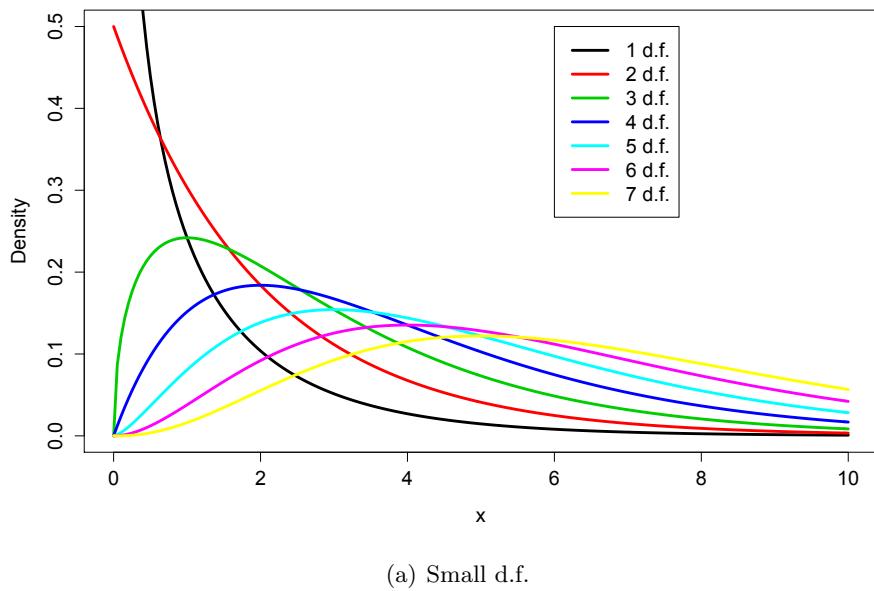
Instead, we rely on a traditional compromise approach, taking advantage of the fact that the most common use of the tables is to find the critical value for hypothesis testing at one of a few levels, such as 0.05 and 0.01.

### Example 9.2: Using the $\chi^2$ table

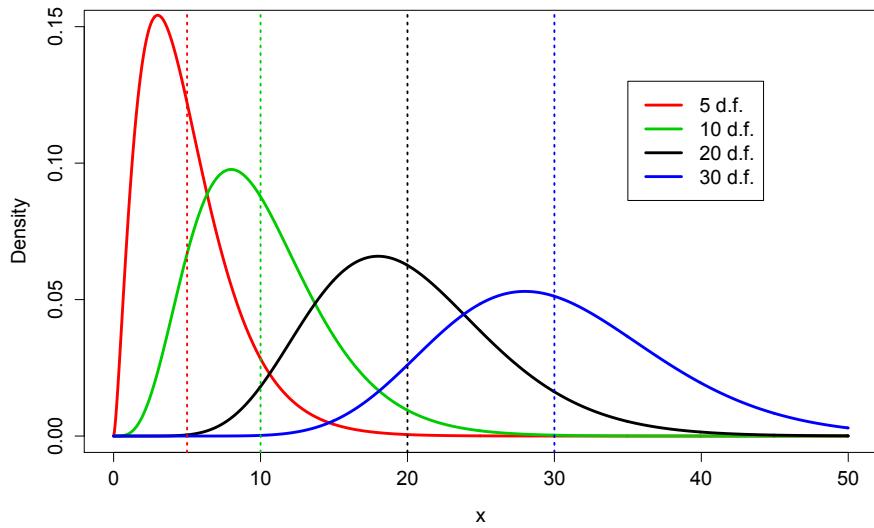
We perform a  $\chi^2$  test to test a certain null hypothesis at the 0.05 significance level, and find an observed  $X^2$  of 12.1 with 5 degrees of freedom. We look on the table and see in the row for 5 d.f. that the critical value is 11.07. Since our observed value is above this, we reject the null hypothesis. On the other hand, we would have retained the null hypothesis had we been testing at the 0.01 level, since the critical value at level 0.01 is 15.09. (The actual p-value is 0.0334.) We show, in figure 9.2, the region corresponding to the 0.01 significance level in green, and the remainder of the 0.05 critical region in red. We have observed 12.1, so the logic tells us that either the data did not come from the null hypothesis, or we happen, purely by chance, to have made a pick that wound up in the tiny red region.

---

<sup>1</sup>It happens to be the same as the distribution of the sum of  $d$  independent random variables, each of which is the square of a standard normal. In particular, the  $\chi^2$  with 1 degree of freedom is the same as the square of a standard normal random variable.



(a) Small d.f.



(b) Moderate d.f.

Figure 9.1: The density of  $\chi^2$  distributions with various degrees of freedom.

d.f.	$P = 0.05$	$P = 0.01$	d.f.	$P = 0.05$	$P = 0.01$
1	3.84	6.63	17	27.59	33.41
2	5.99	9.21	18	28.87	34.81
3	7.81	11.34	19	30.14	36.19
4	9.49	13.28	20	31.41	37.57
5	11.07	15.09	21	32.67	38.93
6	12.59	16.81	22	33.92	40.29
7	14.07	18.48	23	35.17	41.64
8	15.51	20.09	24	36.42	42.98
9	16.92	21.67	25	37.65	44.31
10	18.31	23.21	26	38.89	45.64
11	19.68	24.72	27	40.11	46.96
12	21.03	26.22	28	41.34	48.28
13	22.36	27.69	29	42.56	49.59
14	23.68	29.14	30	43.77	50.89
15	25.00	30.58	40	55.76	63.69
16	26.30	32.00	60	79.08	88.38

Table 9.2: A  $\chi^2$  table

■

### 9.2.2 Large d.f.

Table 9.2 only gives values for up to 60 degrees of freedom. What do we do when the problem gives us more than 60? Looking at Figure 9.1(b) you probably are not surprised to hear that the  $\chi^2$  distribution gets ever closer to a normal distribution when the number of degrees of freedom gets large. Which normal distribution? We already know the mean and variance of the  $\chi^2$ . For large  $d$ , then,

$$\chi^2(d) \text{ is approximately the same as } \mathcal{N}(d, 2d).$$

Thus, the p-value for a given observed  $X^2$  is found by taking

$$Z = \frac{X^2 - d}{\sqrt{2d}},$$

and looking it up on the standard normal table. Conversely, if we want to know the critical value for rejecting  $X^2$  with  $d$  degrees of freedom, with  $d$

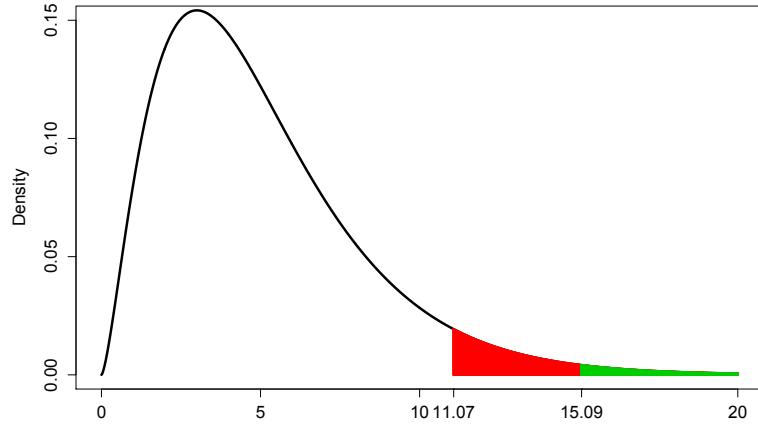


Figure 9.2:  $\chi^2$  density with 5 degrees of freedom. The green region represents 1% of the total area. The red region represents a further 4% of the area, so that the tail above 11.07 is 5% in total.

large, at significance level  $\alpha$ , we start by finding the appropriate  $z$  for the level (if we were doing a Z test). The critical value for  $X^2$  is then

$$\sqrt{2d}z + d.$$

For example, if we were testing at the 0.01 level, with 60 d.f., we would first look for 9950 on the standard normal table, finding that this corresponds to  $z = 2.58$ . (Remember that in a two-tailed test at the 0.01 level, the probability above  $z$  is 0.005.) We conclude that the critical value for  $\chi^2$  with 60 d.f. at the 0.01 level is about

$$2.58\sqrt{120} + 60 = 88.26.$$

The exact value, given on the table, is 88.38. For larger values of d.f. we simply rely on the approximation.

### 9.3 Fixed distributions

We start with a simple example. Suppose we have a six-sided die, and we are wondering whether it is fair — that is, whether each side is equally likely

to come up. We roll it 60 times, and tabulate the number of times each side comes up. The results are given in Table 9.3.

Table 9.3: Outcome of 60 die rolls

Side	1	2	3	4	5	6
Observed Frequency	16	15	4	6	14	5
Expected Frequency	10	10	10	10	10	10

It certainly appears that sides 1,2, and 5 come up more often than they should, and sides 3, 4, and 6 less frequently. On the other hand, some deviation is expected, due to chance. Are the deviations we see here too extreme to be attributed to chance?

Suppose we wish to test the null hypothesis

$$H_0 : \text{Each side comes up with probability } 1/6$$

at the 0.01 significance level. In addition to the Observed Frequency of each side, we have also indicated, in the last row of Table 9.3, the Expected Frequency, which is the probability (according to  $H_0$ ) of the side coming up, multiplied by the number of trials, which is 60. Since each side has probability  $1/6$  (under the null hypothesis), the expected frequencies are all 10. (Note that the observed and expected frequencies both add up to exactly the number of trials.)

We now plug these numbers into our formula for the chi-squared statistic:

$$\begin{aligned} X^2 &= \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \\ &= \frac{(16 - 10)^2}{10} + \frac{(15 - 10)^2}{10} + \frac{(4 - 10)^2}{10} + \frac{(6 - 10)^2}{10} + \frac{(14 - 10)^2}{10} + \frac{(5 - 10)^2}{10} \\ &= 3.6 + 2.5 + 3.6 + 1.6 + 1.6 + 2.5 \\ &= 15.4. \end{aligned}$$

Now we need to decide whether this number is a big one, by comparing it to the appropriate  $\chi^2$  distribution. Which one is the appropriate one? When testing observations against a single fixed distribution, the number of degrees of freedom is always one fewer than the number of categories.<sup>2</sup> There are six categories, so five degrees of freedom.

---

<sup>2</sup>Why minus one? You can think of degrees of freedom as saying, how many different

### Example 9.3: Suicides, continued

We complete Example 9.1. We want to test whether the observed data in Table 9.1 could plausibly have come from the null hypothesis, with the probability for each month given in column 2. In Table 9.4 we add columns for “expected” numbers of suicides. For example, if we look in the column “Female/Exp”, and the row January, we find the number 493. This is obtained from multiplying 5801 (the total number of women in the study) by 0.0849 (the probability of a woman having been born in January, under the null hypothesis). (The numbers in the “expected” columns don’t add up to exactly the same as those in the corresponding “observed” column, because of rounding.)

Month	Prob.	Female		Male		Combined	
		Obs	Exp	Obs	Exp	Obs	Exp
Jan	0.0849	527	493	1774	1790	2301	2283
Feb	0.0773	435	448	1639	1630	2074	2078
Mar	0.0849	454	493	1939	1790	2393	2283
Apr	0.0821	493	476	1777	1731	2270	2207
May	0.0849	535	493	1969	1790	2504	2283
Jun	0.0821	515	476	1739	1731	2254	2207
Jul	0.0849	490	493	1872	1790	2362	2283
Aug	0.0849	489	493	1833	1790	2322	2283
Sep	0.0821	476	476	1624	1731	2100	2207
Oct	0.0849	474	493	1661	1790	2135	2283
Nov	0.0821	442	476	1568	1731	2010	2207
Dec	0.0849	471	493	1690	1790	2161	2283
Total	1.0000	5801	5803	21085	21084	26886	26887

Table 9.4: Suicides 1979–2001 by month of birth in England and Wales 1955–66, taken from [SCB06].

We now test the null hypothesis for the combined sample of men and women, at the 0.01 significance level. We plug these columns

numbers did we really observe? We observed six numbers (the frequency counts for the six sides), but they had to add up to 60, so any five of them determine the sixth one. So we really only observed five numbers.

into equation (9.1), obtaining

$$\begin{aligned} X^2 &= \frac{(2301 - 2283)^2}{2283} + \frac{(2073 - 2078)^2}{2078} \\ &\quad + \frac{(2393 - 2283)^2}{2283} + \dots + \frac{(2161 - 2283)^2}{2283} \\ &= 71.9. \end{aligned}$$

Since there are 12 categories, the number of degrees of freedom is  $12 - 1 = 11$ . We look on the  $\chi^2$  table in row 11 d.f., column  $P = 0.01$ , and find the critical value is 24.73. Since the observed value is higher, we reject the null hypothesis, and conclude that the difference between the observed distribution of suicides' birthmonths and what would be expected purely by chance is highly significant. (The actual p-value is less than  $10^{-10}$ , so there is less than one chance in ten billion that we would observe such a deviation purely by chance, if the null hypothesis were true.)

If we considered only the female data, on the other hand, the  $\chi^2$  value that we would compute from those two columns is 17.4, which is below the critical value. Considered on their own, the female data do not provide strong evidence that suicides' birth months differ in distribution from what would be expected by chance. ■

#### Example 9.4: Distribution of births

Did we use the right null hypothesis in Example 9.1? We assumed that all birthdates should be equally likely.

At [http://www.statistics.gov.uk/downloads/theme\\_population/FM1\\_32/FM1no32.pdf](http://www.statistics.gov.uk/downloads/theme_population/FM1_32/FM1no32.pdf) we have official UK government statistics for births in England and Wales. Table 2.4 gives the data for birth months, and we repeat them in the column “observed” in Table 9.5. The “Expected” column tells us how many births would have been expected in that month under the null hypothesis that all dates are equally likely. We compute  $X^2 = \sum \frac{(Obs_i - Exp_i)^2}{Exp_i} = 477$ . The critical value is still the same as before, which is 24.73, so this is much bigger. (In fact, the p-value

is around  $10^{-16}$ .) So births are definitely not evenly distributed through the year.

Month	Prob.	Observed	Expected
Jan	0.0849	29415	30936
Feb	0.0767	26890	27942
Mar	0.0849	30122	30936
Apr	0.0822	30284	29938
May	0.0849	31516	30936
Jun	0.0822	30571	29938
Jul	0.0849	32678	30936
Aug	0.0849	31008	30936
Sep	0.0822	31557	29938
Oct	0.0849	31659	30936
Nov	0.0822	29358	29938
Dec	0.0849	29186	30936
Total	1.0000	364244	364246

Table 9.5: Observed frequency of birth months, England and Wales, 1993.

We might then decide to try testing a new null hypothesis

$H_0$ : Suicides have the same distribution of birth month as the rest of the population.

In Table 9.6 we show the corresponding calculations. In the column “Prob.” we give the observed empirical fraction of births for that month, as tabulated in Table 9.5. Thus, the probability for January is 0.0808, which is 29415/364244. The “Observed” column copies the final observed column from Table 9.4, and the “Expected” column is obtained by multiplying the “Prob” column by 26886, the total number of suicides in the sample. Using these two columns, we compute  $X^2 = 91.8$ , which is even larger than the value computed before. Thus, when we change to this more appropriate null hypothesis, the evidence that something interesting is going on becomes even stronger.

■

Month	Prob.	Observed	Expected
Jan	0.0808	2301	2171
Feb	0.0738	2074	1985
Mar	0.0827	2393	2223
Apr	0.0831	2270	2235
May	0.0865	2504	2326
Jun	0.0839	2254	2257
Jul	0.0897	2362	2412
Aug	0.0851	2322	2289
Sep	0.0866	2100	2329
Oct	0.0869	2135	2337
Nov	0.0806	2010	2167
Dec	0.0801	2161	2154
Total	1.0000	26886	26885

Table 9.6: Birth months of suicides compared with observed frequencies of birth months in England and Wales, from Table 9.5.

## 9.4 Families of distributions

In many situations, it is not that we want to know whether the data came from a single distribution, but whether it may have come from any one of a whole family of distributions. For example, in Lecture 5 we considered several examples of data that might be Poisson distributed, and for which the failure of the Poisson hypothesis would have serious scientific significance. In such a situation, we modify our  $\chi^2$  hypothesis test procedure slightly:

- (1). Estimate the parameters in the distribution. Now we have a **particular** distribution to represent the null hypothesis.
- (2). Compute the expected occupancy in each cell as before.
- (3). Using these expected numbers and the observed numbers (the data) compute the  $\chi^2$  statistic.
- (4). Compare the computed statistic to the critical value. Important: The degrees of freedom are reduced by one for each parameter that has been estimated.

Thus, if we estimated a single parameter (e.g., Poisson distribution) we look in the row for  $\#$  categories–2 d.f. If we estimated two parameters (e.g., normal distribution) we look in the row for  $\#$  categories–3 d.f.

### 9.4.1 The Poisson Distribution

Consider Example 5.7. We argued there that the distribution of Guillain-Barré Syndrome (GBS) cases by weeks should have been Poisson distributed if the flu vaccine was not responsible for the disease. The data are given in Table 5.3, showing the number of weeks in which different numbers of cases were observed.

If some Poisson distribution were correct, what would be the parameter? We estimated that  $\lambda$ , which is the expected number of cases per week, should be estimated by the observed *average* number of cases per week, which is  $40/30 = 1.33$ . We computed the probabilities for different numbers of cases, assuming a Poisson distribution with parameter 1.33, and multiplied these probabilities by 40 to obtain expected numbers of weeks. We compared the observed to these expected numbers, and expressed the impression that these distributions were different. But the number of observations is small. Could it be that the difference is purely due to chance?

We want to test the null hypothesis  $H_0$ : The data came from a Poisson distribution with a  $\chi^2$  test. We can't use the numbers from Table 5.3 directly, though. As discussed in section 9.2.1, the approximation we use for the  $\chi^2$  distribution depends on the categories all being “large enough”, the rule of thumb being that the expected numbers under the null hypothesis should be at least 5. The last three categories are all too small. So, we group the last four categories together, to obtain the new Table 9.7. The last category includes everything 3 and higher.

Table 9.7: Cases of GBS, by weeks after vaccination

# cases per week	0	1	2	3+
observed frequency	16	7	3	4
probability	0.264	0.352	0.234	0.150
expected frequency	10.6	14.1	9.4	6.0

We now compute

$$X^2 = \frac{(16 - 10.6)^2}{10.6} + \frac{(7 - 14.1)^2}{14.1} + \frac{(3 - 9.4)^2}{9.4} + \frac{(4 - 6.0)^2}{6.0} = 11.35.$$

Suppose we want to test the null hypothesis at the 0.01 significance level. In order to decide on a critical value, we need to know the correct number of degrees of freedom. The reduced Table 9.7 has only 4 categories. There would thus be 3 d.f., were it not for our having estimated a parameter to decide on the distribution. This reduces the d.f. by one, leaving us with 2 degrees of freedom. Looking in the appropriate row, we see that the critical value is 9.21, so we do reject the null hypothesis (the true p-value is 0.0034), and conclude that the data did not come from a Poisson distribution.

#### 9.4.2 The Binomial Distribution

In 1889, A. Geissler published data on births recorded in Saxony over a 10 year period, tabulating the numbers of boys and girls. (The complete data set, together with analysis and interpretation, may be found in [Edw58].) Table 9.8 shows the number of girls in the 6115 families with 12 children. If the gender of successive children are independent, and the probabilities remain constant over time, the number of girls born to a particular family of 12 children should be a binomial random variable with 12 trials and an unknown probability  $p$  of success.

Table 9.8: Numbers of girls in families with 12 children, from Geissler.

# Girls	0	1	2	3	4	5	6
Frequency	7	45	181	478	829	1112	1343
Expected	2.3	26.1	132.8	410.0	854.2	1265.6	1367.3
Probability	0.0004	0.0043	0.0217	0.0670	0.1397	0.2070	0.2236
# Girls	7	8	9	10	11	12	
Frequency	1033	670	286	104	24	3	
Expected	1085.2	628.1	258.5	71.8	12.1	0.9	
Probability	0.1775	0.1027	0.0423	0.0117	0.0020	0.0002	

We can use a Chi-squared test to test the hypothesis that the data follow a Binomial distribution.

$H_0$  : The data follow a Binomial distribution  
 $H_1$  : The data *do not* follow a Binomial distribution

At this point we also decide upon a 0.05 significance level.

From the data we know that  $n = 6115$  and we can estimate  $p$  as

$$\hat{p} = \frac{\bar{x}}{12} = \frac{7(0) + 45(1) + \dots + 3(12)}{12 \times 6115} = 0.4808$$

Thus we can fit a  $\text{Bin}(12, 0.4808)$  distribution to the data to obtain the expected frequencies (E) alongside the observed frequencies (O). The probabilities are shown at the bottom of Table 9.8, and the expectations are found by multiplying the probabilities by 6115. The first and last categories have expectations smaller than 5, so we absorb them into the next categories, yielding Table 9.9.

Table 9.9: Modified version of Table 9.8, with small categories grouped together.

# Girls	0, 1	2	3	4	5	6	7	8	9	10	11, 12
Frequency	52	181	478	829	1112	1343	1033	670	286	104	27
Expected	28.4	132.8	410.0	854.2	1265.6	1367.3	1085.2	628.1	258.5	71.8	13.0
Probability	0.0047	0.0217	0.0670	0.1397	0.2070	0.2236	0.1775	0.1027	0.0423	0.0117	0.0022

The test statistic can then be calculated as

$$\begin{aligned} X^2 &= \frac{(52 - 28.4)^2}{28.4} + \dots + \frac{(27 - 13.0)^2}{13.0} \\ &= 105.95 \end{aligned}$$

The degrees of freedom are given by

$$df = (k - 1) - p = (11 - 1) - 1 = 9$$

Thus, the Critical Region for the test is  $X^2 > 16.92$ .

The test statistics lies well within the Critical Region so we conclude that there is significant evidence against the null hypothesis at the 5% level. We conclude that the sex of newborns in families with 12 children is NOT binomially distributed.

### Explaining the Geissler data

#### Non-examinable. Just if you're interested.

So what is going on? A good discussion may be found in [LA98]. A brief summary is that two things happened:

- (1). Large numbers of children tended to appear in families which started out unbalanced, particularly if they had all girls. That is, a family with three girls would be more likely to have another child than a family with two boys and a girl.
- (2). The  $p$  value really doesn't seem to be the same between families. Some families have a tendency to produce more boys, others more girls.

Note that (1) is consistent with our original hypothesis, that babies all have the same probability  $p$  of being female: We have just pushed the variability from small families to large ones. Think of it this way: Suppose there were a rule that said: Stop when you have 3 children, unless the children are all boys or all girls. Otherwise, keep trying to get a balanced family. Then the small families would be more balanced than you would have expected, and the big families more unbalanced — for instance, half of the four-child families would have all boys or all girls. Of course, it's more complicated than that: Different parents have different ideas about the “ideal” family. But this effect does seem to explain some of the deviation from the binomial distribution.

The statistical analysis in [LA98] tries to pull these effects apart (and also take into account the small effect of identical twins), finding that there is an SD of about 0.16 in the value of  $p$ , the probability of a girl, and furthermore that there is some evidence that some parents produce nothing but girls, or at most have a very small probability of producing boys.

### The Normal Distribution

The following table gives the heights in cm of 100 students. In such a situation we might be interested in testing whether the data follow a Normal distribution or not.

---

Height (cm)	155-160	161-166	167-172	173-178	179-184	185-190
Frequency	5	17	38	25	9	6

---

We can use a Chi-squared test to test the hypothesis that the data follow a Normal distribution.

$H_0$  : The data follow a Normal distribution

$H_1$  : The data *do not* follow a Normal distribution

At this point we also decide upon a 0.05 significance level.

From the data we can estimate the mean and standard deviation using the sample mean and standard deviation

$$\begin{aligned}\bar{x} &= 172 \\ s &= 7.15\end{aligned}$$

To fit a Normal distribution with this mean and variance we need to calculate the probability of each interval. This is done in four straightforward steps

- (i). Calculate the upper end point of each interval ( $u$ )
- (ii). Standardize the upper end points ( $z$ )
- (iii). Calculate the probability  $P(Z < z)$
- (iv). Calculate the probability of each interval
- (v). Calculate the expected cell counts

Height (cm)	155-160	161-166	167-172	173-178	179-184	185-190
Endpoint ( $u$ )	160.5	166.5	172.5	178.5	184.5	$\infty$
Standardized ( $z$ )	-1.61	-0.77	0.07	0.91	1.75	$\infty$
$P(Z < z)$	0.054	0.221	0.528	0.818	0.960	1.00
$P(a < Z < b)$	0.054	0.167	0.307	0.290	0.142	0.040
Expected	5.4	16.7	30.7	29.0	14.2	4.0
Observed	5	17	38	25	9	6

From this table we see that there is one cell with an expected count less than 5 so we group it together with the nearest cell. (A single cell with expected count is on the borderline; we could just leave it. We certainly don't want any cell with expected count less than about 2, and not more than one expected count under 5.)

	Height (cm)				
	155-160	161-166	167-172	173-178	179-190
Expected	5.4	16.7	30.7	29.0	18.2
Observed	5	17	38	25	15

We can then calculate the test statistic as

$$\begin{aligned} X^2 &= \frac{(5.4 - 6)^2}{5.4} + \dots + \frac{(15 - 18.2)^2}{18.2} \\ &= 2.88 \end{aligned}$$

The degrees of freedom are given by

$$df = (\# \text{ categories} - 1) - \# \text{ parameters estimated} = (5 - 1) - 2 = 2.$$

Thus, the Critical Region for the test is  $X^2 > 5.99$ .

The test statistics lies outside the Critical Region so we conclude that the evidence against the null hypothesis is not significant at the 0.05 level.

## 9.5 Chi-squared Tests of Association

This section develops a Chi-squared test that is very similar to the one of the preceding section but aimed at answering a slightly different question. To illustrate the test we use an example, which we borrow from [FPP98, section 28.2].

Table 9.10 gives data on Americans aged 25–34 from the NHANES survey, a random sample of Americans. Among other questions, individuals were asked their age, sex, and handedness.

Table 9.10: NHANES handedness data for Americans aged 25–34

	Men	Women
right-handed	934	1070
left-handed	113	92
ambidextrous	20	8

Looking at the data, it looks as though the women are more likely to be right-handed. Someone might come along and say: “The left cerebral hemisphere controls the right side of the body, as well as rational thought. This proves that women are more rational than men.” Someone else might say, “This shows that women are under more pressure to conform to society’s expectations of normality.” But before we consider this observation as evidence of anything important, we have to pose the question: Does this

reflect a difference in the underlying population, or could it merely be a random effect of sampling?

The research hypothesis in this situation is whether a person's handedness is associated with their sex. In this situation, the null hypothesis would be that there is no association between the two variables. In other words, the null hypothesis is that the two variables are independent.

$H_0$  : The two variables are independent.

$H_1$  : The two variables are associated.

Or, to put it differently, each person gets placed in one of the six cells of the table. The null hypothesis says that which row you're in is independent of the column. This is a lot like the problems we had in section 9.4: Here the family of distributions we're interested in is all the distributions in which the rows are independent of the columns. The procedure is essentially the same:

- (1). Estimate the parameters in the null distribution. This means we estimate the probability of being in each row and each column.
- (2). Compute the expected occupancy in each cell: This is the number of observations times the row probability times the column probability.
- (3). Using these expected numbers and the observed numbers (the data) compute the  $\chi^2$  statistic.
- (4). Compare the computed statistic to the critical value. The number of degrees of freedom is  $(r - 1)(c - 1)$ , where  $r$  is the number of rows and  $c$  the number of columns. (Why? The number of cells is  $rc$ . We estimated  $r - 1$  parameters to determine the row probabilities and  $c - 1$  parameters to determine the column probabilities. So we have  $r + c - 2$  parameters in all. By our standard formula,

$$\text{d.f.} = rc - 1 - (r + c - 2) = (r - 1)(c - 1).$$

We wish to test the null hypothesis at the 0.01 significance level. We extend the table to show the row and column fractions in Table 9.11. Thus, we see that 89.6% of the sample were right-handed, and 47.7% were male. The fraction that were right-handed males would be  $0.896 \times 0.477 = 0.427$  under the null hypothesis. Multiplying this by 2237, the total number of observations, we obtain 956, the expected number of right-handed men under the

null hypothesis. We repeat this computation for all six categories, obtaining the results in Table 9.12. (Note that the row totals and the column totals are identical to the original data.) We now compute the  $\chi^2$  statistic, taking the six “observed” counts from the black Table 9.10, and the six “expected” counts from the red Table 9.12:

	Men	Women	Total	Fraction
right-handed	934	1070	2004	0.896
left-handed	113	92	205	0.092
ambidextrous	20	8	28	0.013
Total	1067	1170	2237	
Fraction	0.477	0.523		

Table 9.11: NHANES handedness data for Americans aged 25–34.

	Men	Women
right-handed	956	1048
left-handed	98	107
ambidextrous	13	15

Table 9.12: Expected counts for NHANES handedness data, computed from fractions in Table 9.11.

$$X^2 = \frac{(934 - 956)^2}{956} + \frac{(1070 - 1048)^2}{1048} + \dots + \frac{(8 - 15)^2}{15} = 12.4.$$

The degrees of freedom are  $(3-1)(2-1) = 2$ , so we see on Table 9.2 that the critical value is 9.21. Since the observed value is higher, we reject the null hypothesis, and conclude that the difference in handedness between men and women is not purely due to chance.

Remember, this does not say anything about the reason for the difference! It may be something interesting (e.g., women are more rational) or it may be something dull (e.g., men were more likely to think the interviewer would be impressed if they said they were left-handed). All the hypothesis

test tells us is that we would most likely have found a difference in handedness between men and women if we had surveyed the whole population.

# Lecture 10

## The T distribution and Introduction to Sampling

### 10.1 Using the T distribution

You may have noticed a hole in the reasoning we used in reasoning about the husbands' heights in section 7.1. Our computations depended on the fact that

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has standard normal distribution, when  $\mu$  is the population mean and  $\sigma$  is the population standard deviation. But *we don't know  $\sigma$ !* We did a little slight of hand, and substituted the sample standard deviation

$$S := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

for the (unknown) value of  $\sigma$ . But  $S$  is only an estimate for  $\sigma$ : it's a random variable that might be too high, and might be too low. So, what we were calling  $Z$  is not really  $Z$ , but a quantity that we should give another name to:

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}}. \quad (10.1)$$

If  $S$  is too big, then  $Z > T$ , and if  $S$  is too small then  $Z < T$ . On average, you might suppose,  $Z$  and  $T$  would be about the same — and, in this you would be right. Does the distinction matter then?

Since  $T$  has an extra source of error in the denominator, you would expect it to be more widely scattered than  $Z$ . That means that if you compute  $T$  from the data, but look it up on a table computed from the distribution of  $Z$  — the standard normal distribution — you would underestimate the probability of a large value. The probability of rejecting a true null hypothesis (Type I error) will be larger than you thought it was, and the confidence intervals that you compute will be too narrow. This is very bad! If we make an error, we always want it to be on the side of underestimating our confidence.

Fortunately, we can compute the distribution of  $T$  (sometimes called “Student’s t”, after the pseudonym under which statistician William Gossett published his first paper on the subject, in 1908). While the mathematics behind this is beyond the scope of this course, the results can be found in tables. These are a bit more complicated than the normal tables, because there is an extra parameter: Not surprisingly, the distribution depends on the number of samples. When the estimate is based on very few samples (so that the estimate of SD is particularly uncertain) we have a distribution which is far more spread out than the normal. When the number of samples is very large, the estimate  $s$  varies hardly at all from  $\sigma$ , and the corresponding t distribution is very close to normal. As with the  $\chi^2$  distribution, this parameter is called “degrees of freedom”. For the  $T$  statistic, the number of degrees of freedom is just  $n - 1$ , where  $n$  is the number of samples being averaged. Figure 10.1 shows the density of the t distribution for different degrees of freedom, together with that of the normal. Note that the t distribution is symmetric around 0, just like the normal distribution.

Table 10.1 gives the critical values for a level 0.05 hypothesis test when  $Z$  is replaced by  $t$  with different numbers of degrees of freedom. In other words, if we define  $t_\alpha(d)$  to be the number such that  $\mathbb{P}\{T < t_\alpha\} = \alpha$  when  $T$  has the Student distribution with  $d$  degrees of freedom, Table 10.1(a) gives values of  $t_{0.95}$ , and Table 10.1(b) gives values of  $t_{0.975}$ . Note that the values of  $t_\alpha(d)$  decrease as  $d$  increases, approaching a maximum, which is  $z_\alpha = t_\alpha(\infty)$ .

### 10.1.1 Using t for confidence intervals: Single sample

Suppose we have observations  $x_1, \dots, x_n$  from a normal distribution, where the mean and the SD are both unknown. To compute confidence intervals

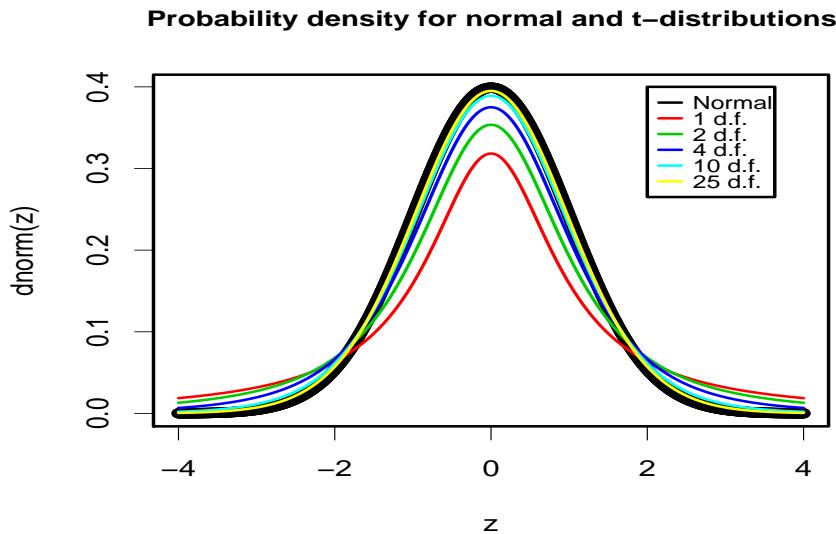


Figure 10.1: The standard normal density together with densities for the t distribution with different degrees of freedom.

Table 10.1: Cutoffs for hypothesis tests at the 0.95 level, using the t statistic with different degrees of freedom. The  $\infty$  level is the limit for a very large number of degrees of freedom, which is identical to the distribution of the Z statistic.

(a) one-tailed test		(b) two-tailed test	
degrees of freedom	critical value	degrees of freedom	critical value
1	6.31	1	12.7
2	2.92	2	4.30
4	2.13	4	2.78
10	1.81	10	2.23
50	1.68	50	2.01
$\infty$	1.64	$\infty$	1.96

with the t statistic, we follow the same procedures as in section 7.1, substituting  $s$  for  $\sigma$ , and the quantiles of the t distribution for the quantiles of the normal distribution: that is, where we looked up a number  $z$  on the normal table, such that  $P(Z < z)$  was a certain probability, we substitute a number  $t$  such that  $P(T < t)$  is that same probability, where  $T$  has the Student T distribution with the right number of degrees of freedom. Thus, if we want a 95% confidence interval, we take

$$\bar{X} \pm t \times \frac{s}{\sqrt{n}},$$

where  $t$  is found in the column marked “ $P = 0.05$ ” on the T-distribution table — 0.05 being the probability above  $t$  that we are excluding. It corresponds, of course, to  $P(T < t) = 0.975$ .

### Example 10.1: Heights of British men

In section 7.1 we computed a confidence interval for the heights of married British men, based on a sample of size 198. Since we were using the sample SD to estimate the population SD, we should have used the t quantiles with 197 degrees of freedom, rather than the Z quantiles. If you look on a table of the t distribution you won’t find a row corresponding to 197 degrees of freedom, though. Why not? The t distribution with 197 degrees of freedom is almost indistinguishable from the normal distribution. To give an example, the multiplier for a symmetric 90% normal confidence interval is  $z = 1.645$ ; the corresponding t quantile is  $t(197) = 1.653$ , so the difference is less than 1%. There is no real application where you are likely to be able to notice an error of that magnitude. ■

### Example 10.2: Kidney dialysis

A researcher measured the blood level of phosphate in the blood of dialysis patients on six consecutive clinical visits.<sup>1</sup> It is important to maintain the levels of various nutrients in appropriate bounds during dialysis treatment. The values are known to vary

---

<sup>1</sup>This example is adapted from [MM98, p.529], where it was based on a Master’s thesis of Joan M. Susic at Purdue University.

approximately according to a normal distribution. For one patient, the values (in mg/dl) were measured 5.6, 5.1, 4.6, 4.8, 5.7, 6.4. What is a symmetric 99% confidence interval for the patient's true phosphate level?

We compute

$$\begin{aligned}\bar{X} &= \frac{1}{6}(5.6 + 5.1 + 4.6 + 4.8 + 5.7 + 6.4) = 5.4 \text{mg/dl} \\ s &= \sqrt{\frac{1}{5} \left( (5.6 - 5.4)^2 + (5.1 - 5.4)^2 + (4.6 - 5.4)^2 + (4.8 - 5.4)^2 + (5.7 - 5.4)^2 + (6.4 - 5.4)^2 \right)} \\ &= 0.67 \text{mg/dl.}\end{aligned}$$

The number of degrees of freedom is 5. Thus, the symmetric confidence interval will be

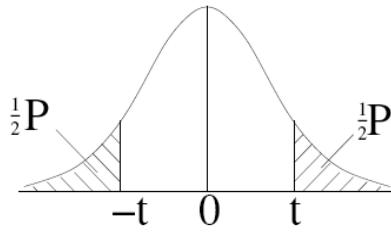
$$\left( 5.4 - t \frac{0.67}{\sqrt{6}}, 5.4 + t \frac{0.67}{\sqrt{6}} \right) \text{mg/dl},$$

where  $t$  is chosen so that the T variable with 5 degrees of freedom has probability 0.01 of being bigger than  $t$ . ■

### 10.1.2 Using the T table

T tables are like the  $\chi^2$  table. For Z, the table in your official booklet allows you to choose your value of Z, and gives you the probability of finding Z below this value. Thus, if you were interested in finding  $z_\alpha$ , you would look to find  $\alpha$  inside the table, and then check which index Z corresponds to it. In principle, we could have a similar series of T tables, one for each number of degrees of freedom. To save space, though, and because people are usually not interested in the entire t distribution, but only in certain cutoffs, the T tables give much more restricted information. The rows of the T table represent degrees of freedom, and the columns represent cutoff probabilities. The values in the table are then the values of t that give the cutoffs at those probabilities. One peculiarity of these tables is that, whereas the Z table gives one-sided probabilities, the t table gives two-sided probabilities. This makes things a bit easier when you are computing symmetric confidence intervals, which is all that we will do here.

The probability we are looking for is 0.01, which is the last column of the table, so looking in the row for 5 d.f. (see Figure 10.2) we see that



Probability  $P$  of lying outside  $\pm t$

d.f.	P=0.10	P=0.05	P=0.02	P=0.01
1	6.31	12.71	31.82	63.7
2	2.92	4.30	6.96	9.93
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71

Figure 10.2: Excerpt from the official t table, p. 21.

the appropriate value of  $t$  is 4.03. Thus, we can be 99% confident that the patient's true average phosphate level is between 4.3mg/dl and 6.5mg/dl. Note that if we had known the SD for the measurements to be 0.67, instead of having estimated it from the observations, we would have used  $z = 2.6$  (corresponding to a one-sided probability of 0.995) in place of  $t = 4.03$ , yielding a much narrower confidence interval.

### Summary

If you want to compute an  $\alpha \times 100\%$  confidence interval for the population mean of a normally distributed population based on  $n$  samples you do the following:

- (1). Compute the sample mean  $\bar{x}$ .
- (2). Compute the sample SD  $s$ .
- (3). Look on the table to find the number  $t$  in the row corresponding to  $n - 1$  degrees of freedom and the column corresponding to  $\alpha$ .
- (4). The confidence interval is from  $\bar{x} - st/\sqrt{n}$  to  $\bar{x} + st/\sqrt{n}$ . In other words, we are  $\alpha \times 100\%$  confident that  $\mu$  is in this range.

### 10.1.3 Using t for Hypothesis tests

We continue Example 10.2. Suppose 4.0 mg/dl is a dangerous level of phosphate, and we want to be 99% sure that the patient is, on average, above that level. Of course, all of our measurements are above that level, but they are also quite variable. It could be that all six of our measurements were exceptionally high. How do we make a statistically precise test?

Let  $H_0$  be the null hypothesis, that the patient's phosphate level is actually  $\mu_0 = 4.0\text{mg/dl}$ . The alternative hypothesis is that it is a different value, so this is a two-sided test. Suppose we want to test, at the 0.01 level, whether the null hypothesis could be consistent with the observations. We compute

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = 5.15.$$

This statistic  $T$  has the t distribution with 5 degrees of freedom. The critical value is the value  $t$  such that the probability of  $|T|$  being bigger than  $t$  is 0.01. This is the same value that we looked up in Example 10.2, which is 4.03. Since our T value is 5.15, we reject the null hypothesis. That is,  $T$  is much too big: the probability of such a high value is smaller than 0.01. (In fact, it is about 0.002.) Our conclusion is that the true value of  $\mu$  is not 4.0.

In fact, though, we're likely to be concerned, not with a particular value of  $\mu$ , but just with whether  $\mu$  is too big or too small. Suppose we are concerned to be sure that the average phosphate level  $\mu$  is really *at least*  $\mu_0 = 4.0\text{mg/dl}$ . In this case, we are performing a one-sided test, and we will reject  $T$  values that are too large (meaning that  $\bar{x}$  is too large to have plausibly resulted from sampling a distribution with mean  $\mu_0$ ). The computation of  $T$  proceeds as before, but now we have a different cutoff, corresponding to a probability twice as big as the level of the test, so 0.02. (This is because of the peculiar way the table is set up. We're now only interested in the probability in the upper tail of the t distribution, which is 0.01, but the table is indexed according to the total probability in both tails.) This is  $t = 3.36$ , meaning that we would have been more likely to reject the null hypothesis.

### 10.1.4 When do you use the Z or the T statistics?

When testing or making a confidence interval for the population mean,

- If you know the population variance, use  $Z$ .
- If you estimate the population variance from the sample, use  $T$ .
- Exception: Use  $Z$  when estimating a proportion.

- Another exception: If the number of samples is large there is no difference between Z and t. You may as well use Z, which is conceptually a bit simpler. For most purposes,  $n = 50$  is large enough to do only Z tests.

### 10.1.5 Why do we divide by $n - 1$ in computing the sample SD?

*This section is not examinable.* The population variance is defined to be the average of the squared deviations from the mean (and the SD is defined to be the square root of that):

$$\sigma_x^2 = \text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Why is it, then, that we estimate variance and SD by using a sample variance and sample SD in which  $n$  in the denominator is replaced by  $n - 1$ ?:

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

The answer is that, if  $x_1, \dots, x_n$  are random samples from a distribution with variance  $\sigma^2$ , then  $s_x^2$  is a better estimate for  $\sigma^2$  than is  $\sigma_x^2$ . Better in what sense? The technical word is “unbiased,” which simply means that over many trials it will turn out to be correct. In other words,  $\sigma_x^2$  is, on average a bit too small, by exactly a factor of  $(n - 1)/n$ . It makes sense to expect it to be too small, since you would expect

$$\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

to be just right, on average, if only we knew what  $\mu$  was. Replacing  $\mu$  by the estimate  $\bar{x}$  will make it smaller. (In fact, for any numbers  $x_1, \dots, x_n$ , the number  $a$  that makes  $\sum(x_i - a)^2$  as small as possible is  $a = \bar{x}$ . Can you see why?)

As an example, consider the case  $n = 2$ , and let  $X_1, X_2$  be two random

choices from the distribution. Then  $\bar{X} = (X_1 + X_2)/2$ , and

$$\begin{aligned}\sigma_X^2 &= \frac{1}{2} ((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2) \\ &= \left( \frac{X_1 - X_2}{2} \right)^2 \\ &= \frac{1}{4} ((X_1 - \mu) + (\mu - X_2))^2 \\ &= \frac{1}{4} [(X_1 - \mu)^2 + (\mu - X_2)^2 + 2(X_1 - \mu)(\mu - X_2)].\end{aligned}$$

How big is this on average? The first two terms in the brackets will average to  $\sigma^2$  (the technical term is, their *expectation* is  $\sigma^2$ ), while the last term averages to 0. The total averages then to just  $\sigma^2/2$ .

## 10.2 Paired-sample t test

A study<sup>2</sup> was carried out to study the effect of cigarette smoking on blood clotting. Some health problems that smokers are prone to are a result of abnormal blood clotting. Blood was drawn from 11 individuals before and after they smoked a cigarette, and researchers measured the percentage of blood platelets — the factors responsible for initiating clot formation — that aggregated when exposed to a certain stimulus. The results are shown in Table 10.2.

We see that the “Before” numbers tend to be larger than the “After” numbers. But could this be simply a result of random variation? After all, there is quite a lot of natural variability in the numbers.

Imagine that we pick a random individual, who has a normally distributed Before score  $X_i$ . Smoking a cigarette adds a random effect (also normally distributed)  $D_i$ , to make the After score  $Y_i$ . It’s a mathematical fact that, if  $X$  and  $D$  are independent, and  $Y = X + D$ , then

$$Var(Y) = Var(X) + Var(D).$$

We are really interested to know whether  $D_i$  is positive on average, which we do by comparing the observed average value of  $d_i$  to the SD of  $d_i$ . But when we did the computation, we did not use the SD of  $d_i$  in the denominator; we used the SD of  $x_i$  and  $y_i$ , which is much bigger. That is, the average difference between Before and After numbers was found to be not large

---

<sup>2</sup>[Lev73], discussed in [Ric95].

Before	After	Difference
25	27	2
25	29	4
27	37	10
44	56	12
30	46	16
67	82	15
53	57	4
53	80	27
52	61	9
60	59	-1
28	43	15

Table 10.2: Percentage of blood platelets that aggregated in 11 different patients, before and after smoking a cigarette.

enough relative to the SD to be statistically significant, but it was *the wrong SD*. Most of the variability that we found was variation between individuals in their Before scores, not variability in the change due to smoking.

We can follow exactly the same procedure as in section 10.1.3, applied now to the differences. We find that the mean of the differences is  $\bar{d} = -10.3$ , and the SD is  $s_d = 7.98$ . The t statistic is

$$T = \frac{-10.3}{7.98/\sqrt{11}} = -4.28.$$

The cutoff for  $p = 0.05$  at 10 d.f. is found from the table to be 2.23, so we can certainly reject the null hypothesis at the 0.05 level. The difference between the Before and After measurements is found to be statistically significant. (In fact, the p-value may be calculated to be about 0.002.)

## 10.3 Introduction to sampling

### 10.3.1 Sampling with and without replacement

It may seem odd that the computations in section 7.1 take no account of the size of the population that we are sampling from. After all, if these 200 men were all the married men in the UK, there would be no sampling error at all. And if the total population were 300, so that we had sampled 2 men out of 3, there would surely be less random error than if there are 20 million

men in total, and we have sampled only 1 man out of 100,000. Indeed, this is true, but the effect vanishes quite quickly as the size of the population grows.

Suppose we have a box with  $N$  cards in it, each of which has a number, and we sample  $n$  cards *without replacement*, drawing numbers  $X_1, \dots, X_n$ . Suppose that the cards in the box were themselves drawn from a normal distribution with variance  $\sigma^2$ , and let  $\mu$  be the population mean — that is, the mean of the numbers in the box. The sample mean  $\bar{X}$  is still normally distributed, with expectation  $\mu$ , so the only question now is to determine the standard error. Call this standard error  $SE_{NR}$  (NR=no replacement), and the SE computed earlier  $SE_{WR}$  (WR=with replacement). It turns out that the standard error is precisely

$$SE_{NR} = SE_{WR} \sqrt{\left(1 - \frac{n-1}{N-1}\right)} = \frac{\sigma}{\sqrt{n}} \sqrt{\left(1 - \frac{n-1}{N-1}\right)}. \quad (10.2)$$

Thus, if we had sampled 199 out of 300, the SE (and hence also the width of all our confidence intervals) would be multiplied by a factor of  $\sqrt{101/299} = 0.58$ , so would be barely half as large. If the whole population is 1000, so that we have sampled 1 out of 5, the correction factor has gone up to 0.89, so the correction is only by about 10%. And if the population is 10,000, the correction factor is 0.99, which is already negligible for nearly all purposes.

Thus, if the 199 married men had been sampled from a town with just 300 married men, the 95% confidence interval for the average height of married men in the town would be  $1732\text{mm} \pm 2 \cdot 0.58 \cdot 4.9\text{mm} = 1732\text{mm} \pm 5.7\text{mm}$ , so about  $(1726, 1738)\text{mm}$ , instead of the 95% confidence interval computed earlier for sampling with replacement, which was  $(1722, 1742)\text{mm}$ .

The size of the sample matters far more than the size of the population (unless you are sampling a large fraction of the population without replacement).

### 10.3.2 Measurement bias

Bias is a crucial piece of the picture. This is the piece of the error that is systematic: For instance, if you are measuring plots of land with a metre stick, some of your measures will by chance be too big, and some will be too small. The random errors will tend to be normally distributed with mean 0. Averaging more measurements produces narrower confidence bounds.

Suppose your metre stick is actually 101 cm long, though. Then all of your measurements will start out about 1% too short before you add random error on to them. Taking more measurements will not get you closer to the true value, but rather to the ideal biased measure. The important lesson is: Statistical analysis helps us to estimate the extent of random error. Bias remains.

Of course, statisticians are very concerned with understanding the sources of bias; but bias is very subject-specific. The bias that comes in conducting a survey is very different from the bias that comes from measuring the speed of blink reflexes in a psychology experiment.

Better measurement procedures, and better sampling procedures, can reduce bias. Increasing numbers of measurements, or larger samples, reduce the random error. Both cost time and effort. The trick is to find the optimum tradeoff.

### 10.3.3 Bias in surveys

An excellent discussion of the sources of bias in surveys may be found in the book *Statistics*, by Freedman, Pisani, and Purves. Here are some types of bias characteristic of surveys that researchers have given names to:

#### Selection bias

The distribution you sample from may actually differ from the distribution you thought you were sampling from. In the simplest (and most common) type of survey, you mean to be doing a so-called “simple random sample” of the population: Each individual has the same chance of being in the sample. It’s easy to see how this assumption could go wrong. How do you pick a random set of 1000 people from all 60 million people in Britain? Do you dial a random telephone number? But some people have multiple telephone lines, while others have none. And what about mobiles? Some people are home more than others, so are more likely to be available when you call. And so on. All of these factors can bias a survey. If you survey people on the street, you get only the people who are out and about.

Early in the 20th century, it was thought that surveys needed to be huge to be accurate. The larger the better. Then some innovative pollsters, like George Gallup in the US, realised that for a given amount of effort, you would get a better picture of the true population distribution by taking

a smaller sample, but putting more effort into making sure it was a good random sample. More random error, but less bias. The biggest advantage is that you can compute how large the random error is likely to be, whereas bias is almost entirely unknowable.

In section 7.1 we computed confidence intervals for the heights of British men (in 1980) on the basis of 199 samples from the OPCS survey. In fact, the description we gave of the data was somewhat misleading in one respect: The data set we used actually gave the paired heights of husbands and wives. Why does this matter? This sample is potentially biased because the only men included are married. It is not inconceivable that unmarried men have different average height from married men. In fact, the results from the complete OPCS sample are available [RSKG85], and the average height was found to be 1739mm, which is slightly higher than the average height for married men that we found in our selective subsample, but still within the 95% confidence interval.

The most extreme cases of selection bias arise when the sample is **self-selected**. For instance, if you look on a web site for a camera you're interested in, and see that 27 buyers said it was good and 15 said it was bad, what can you infer about the true percentage of buyers who were satisfied? Essentially nothing. We don't know what motivated those particular people to make their comments, or how they relate to the thousands of buyers who didn't comment (or commented on another web site).

### Non-response bias

If you're calling people at home to survey their opinions about something, they might not want to speak to you — and the people who speak to you may be different in important respects from the people who don't speak to you. If you distribute a questionnaire, some will send it back and some won't. Again, the two groups may not be the same.

As an example, consider the health questionnaire that was mailed to a random sample of 6009 residents of Somerset Health District in 1992. The questionnaire consisted of 43 questions covering smoking habits, eating patterns, alcohol use, physical activity, previous medical history and demographic and socio-economic details. 57.6% of the surveys were returned, and on this basis the health authorities could estimate, for instance, that 24.2% of the population were current smokers, or that 44.3% engage in "no moderate or vigorous activity". You might suspect something was wrong when you see that 45% of the respondents were male — as compared with just under 50% of the population of Somerset (known from the census). In

fact, a study [HREG97] evaluated the nonresponse bias by attempting to contact a sample of 437 nonrespondents by telephone, and asking them some of the same questions. 236 were reached and agreed to take part. It turned out that 57.5% of them were male; 32.3% of the contacted nonrespondents were current smokers, and 67.8% of them reported no moderate or vigorous activity. Thus, nonresponse bias had led to substantial underestimation of these important risk factors in the population.

### Response bias

Sometimes subjects don't respond. Sometimes they do, but they don't tell the truth. "Response bias" is the name statisticians give to subjects giving an answer that they think is more acceptable than the true answer. For instance, one 1973 study [LSS73] asked women to express their opinion (from 1=strongly agree to 5=strongly disagree) to "feminist" or "anti-feminist" statements. When the interviewer was a woman, the average response to the statement "The woman's place is in the home" was 3.09 — essentially neutral — but this shifted to clear disagreement (3.80) when the interviewer was a man. Similarly for "Motherhood and a career should not be mixed" (2.96 as against 3.62). On the other hand, those interviewed by women averaged close to neutral (2.78) on the statement "A completely liberalized abortion law is right," whereas those interviewed by men were close to unanimous strong agreement (1.31 average).

On a similar note, in the preceding 2 November presidential election, just over 60% of registered voters cast ballots; when Gallup polled the public less than three weeks later, 80% said they had voted. Another well-known anomaly is the difference between heterosexual men and women in the number of lifetime sexual partners they report (which logically must be the same, on average). [BS]

### Example 10.3: Tsunami donations

Carl Bialik [Bia05] pointed to a 2005 poll by the Gallup organisation, in which Americans were asked whether they had donated to aid victims of the recent Asian tsunami. A stunning 33% said they had. The pollsters went on to ask those who said they had donated how much they had donated. The results, available from the Gallup web site, are given in Table 10.3. Putting aside the two very large donors, we see that 1/4 of households

\$Amount	0	1	5	10	15	20	25	30	35	40	50	55	75
N	3	6	7	19	1	20	22	7	1	4	28	1	1
\$Amount	78	79	80	100	120	150	151	180	200	250	300	325	
N	1	1	0	73	1	4	2	1	15	5	8	0	
\$Amount	400	500	550	750	1000	1100	2000	2500	5000	9999+			
N	1	9	1	0	4	1	1	1	3	2			

Table 10.3: The amounts claimed to have been donated to tsunami relief by 254 respondents to a Gallup survey in January 2005, out of 1008 queried.

donated an average of \$192, meaning \$48 per household. Since there are about 110 million households in the US, that brings us to a total of more than \$5 billion donated. Bialik noted that “the *Chronicle of Philanthropy*, a trade publication, reported a total of \$745 million donated by private sources”, leaving a gap of more than \$4 billion between what people said they donated, and what relief organisations received.

■

### Ascertainment bias

You analyse the data you have, but don’t know which data you never got to observe. This can be a particular problem in a public health context, where you only get reports on the illnesses serious enough to people to seek medical treatment. The recent swine flu outbreak Before the recent outbreak of swine flu, a novel bird flu was making public health experts nervous as it spread through the world from its site of origin in East Asia. While it mainly affects waterfowl, occasional human cases have occurred. Horrific mortality rates, on the order of 50%, have been reported. Thorson *et al.* [TPCE06] pointed out, though, that most people with mild cases of flu never come to the attention of the medical system, particularly in poor rural areas of Vietnam and China, where the disease is most prevalent. They found evidence of a high rate of “flulike illness” associated with contact with poultry among the rural population in Vietnam. Quite likely, then, many of these illnesses were mild cases of the avian influenza. Mortality rate is the probability of cases of disease resulting in death, which we estimate from the fraction of **observed** cases resulting in death. In this case, though, the sample of observed cases was biased: A severe case of flu was more likely to be observed than a

mild case. Thus, while the fraction of observed cases resulting in death was quite high — in Vietnam 2003–5 there were 87 confirmed cases, of which 38 resulted in death — this likely does not reflect accurately the fraction of deaths among all cases in the population. In all likelihood, the 38 includes nearly all the deaths, but the 87 represents only a small fraction of the cases.

During World War II the statistician Abraham Wald worked with the US air force to analyse the data on damage to military airplanes from enemy fire. The question: Given the patterns of damage that we observe, what would be the best place to put extra armour plating to protect the aircraft. (You can't put too much armour on, because it makes the aircraft too heavy.) His answer: Put armour in places where you never see a bullet hole. Why? The bullet holes you see are on the planes that made it back. If you never see a bullet hole in some part of the plane, that's probably because the planes that were hit there didn't make it back. (His answer was more complicated than that, of course, and involved some careful statistical calculations. For a discussion, see [MS84].)

#### 10.3.4 Measurement error

An old joke says, “If you have one watch you know what time it is. If you have two watches you’re never sure.” What do you do when multiple measurements of the same quantity give different answers. One possibility is to try to find out which one is “right”. Anders Hald, in his history of statistics [Hal90], wrote

The crude instruments used by astronomers in antiquity and the Middle Ages could lead to large [...] errors. By planning their observations astronomers tried to balance positive and negative systematic errors. If they made several observations of the same object, they usually selected the best as estimator of the true value, the best being defined from such criteria as the occurrence of good observational conditions, special care having been exerted, and so on.

One of the crucial insights that spurred the development of statistical theory is that the sampling and error problems are connected. Each measurement can be thought of as a sample from the population of possible measurements, and the whole sample tells you more about the population — hence about the hidden “true” value — than any single measurement could.

There are no measurements without error.

Of course, in most real settings measurement error is mixed with population variation. Furthermore, effort put into perfecting the measurement might be more beneficially put into acquiring more samples.

A very basic model is the **Gauss model of errors**, which breaks down the error into two pieces, the **chance error**, which is a random quantity with expectation 0:

$$\text{measurement} = \text{true value} + \text{chance error} + \text{bias}. \quad (10.3)$$

This is a bit of a fiction, whereby we simply define chance error to be zero on average, and call any trend that is left over “bias”.

In a study where you measure each individual once, population variability and random measurement error are mixed up together: If your measurements have a high SD, you don't know whether that is because the population is really variable, or because each measurement had a large error attached to it. You may not care. If you do care, you can make multiple measurements for each individual. Then the methods of chapter 14 will show you how to separate out the two pieces.



# Lecture 11

## Comparing Distributions

### 11.1 Normal confidence interval for difference between two population means

Consider again the sample of 198 men's heights, which we discussed in sections 7.1 and 10.3.3. As mentioned there, the data set gives paired heights of husbands and wives, together with their ages, and the age of the husband at marriage. This might allow us to pose a different sort of question. For instance, What is the average difference in height between men who married early and the men who married late? We summarise the data in Table 11.1, defining "early-married" to mean before age 30.

What does this tell us? We know that the difference in our sample is 19mm, but does this reflect a true difference in the population at large, or could it be a result of mere random selection variation? To put it differently, how sure are we that if we took another sample of 199 and measured them, that we wouldn't find a very different pattern?

	early (< 30)	late ( $\geq 30$ )	unknown	total
number	160	35	3	198
mean	1735	1716	1758	1732
SD	67	78	59	69

Table 11.1: Summary statistics for heights in mm of 198 married men, stratified by age at marriage: early (before age 30), late (age 30 or later), or unknown.

Let  $\mu_X$  be the true average in the population of the heights of early-married men, and  $\mu_Y$  the true average for late-marrieds. The parameter we are interested in is  $\mu_{X-Y} := \mu_X - \mu_Y$ . Obviously the best estimate for  $\mu_{X-Y}$  will be  $\bar{X} - \bar{Y}$ , which will be normally distributed with the right mean, so that a symmetric level confidence interval will be  $(\bar{X} - \bar{Y}) \pm z \times SE$ . But what is the appropriate standard error for the difference?

Since the variance of a sum of independent random variables is the sum of their variances, we see that

$$SE_{X-Y}^2 = Var(\bar{X} - \bar{Y}) = Var(\bar{X}) + Var(\bar{Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y},$$

where  $\sigma_X$  is the standard deviation for the  $X$  variable (the height of early-marrieds) and  $\sigma_Y$  is the standard deviation for the  $Y$  variable (the height of late-marrieds);  $n_X$  and  $n_Y$  are the corresponding numbers of samples. This gives us the standard formula:

$$SE_{\bar{X}-\bar{Y}} = \sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} = \sigma \sqrt{\frac{1}{n_Y} + \frac{1}{n_X}} \text{ if } \sigma_X = \sigma_Y.$$

Formula 11.1: Standard error for the difference between two normally distributed variables

Thus,  $\mu_{X-Y} = \hat{\mu}_{X-Y} + SE_{X-Y} \cdot Z$ , where  $Z$  has standard normal distribution. Suppose now we want to compute a 95% confidence interval for the difference in heights of the early- and late-marrieds. The point estimator we know is +19mm, and the SE is

$$\sqrt{\frac{67^2}{160} + \frac{78^2}{35}} \approx 14\text{mm}.$$

The confidence interval for the difference ranges then from -9mm to +47mm. Thus, while our best guess is that the early-marrieds are on average 19mm taller than the late-marrieds, all that we can say with 95% confidence, on the basis of our sample, is that the difference in height is between -9mm and +47mm. That is, heights are so variable, that a sample of this size might easily be off by 28 mm either way from the true difference in the population.

## 11.2 Z test for the difference between population means

Suppose we wish to make an important argument about the way people choose marriage partners, and base it on the observation that men who marry young tend to be taller — hence that taller men marry younger. But is this true? Or could it be just the particular sample we happened to get, and might we have come to the opposite conclusion from a different sample? One way of answering this is to point out that the 95% confidence interval we computed in section 11.1 includes 0. Another way of expressing exactly the same information is with a significance test.

We have assumed that our samples come from normal distributions, with known (and distinct) variances  $\sigma_X^2$  and  $\sigma_Y^2$ , and unknown (and possibly distinct) means  $\mu_X$  and  $\mu_Y$ . (In fact, the variances have been estimated from the data, but the number of observations is large enough that we can ignore this limitation. For smaller numbers of observations, see sections 11.4 and 11.5.1. The null hypothesis, which says “nothing interesting happened, it’s just chance variation” is

$$H_0 : \mu_X = \mu_Y,$$

The two-sided alternative is  $\mu_X \neq \mu_Y$ .

Using our results from section 11.1 compute the test statistic

$$Z = \frac{\mu_X - \mu_Y}{SE_{X-Y}} = \frac{19mm}{14mm} = 1.4.$$

If we were testing at the 0.05 level, we would not reject the null hypothesis, we would reject values of  $Z$  bigger than 1.96 (in absolute value). Even testing at the 0.10 level we would not reject the null, since the cutoff is 1.6. Our conclusion is that the difference in heights between the early-married and late-married groups is not statistically significant. Notice that this is precisely equivalent to our previous observation that the symmetric 95% confidence interval includes 0.

If we wish to test  $H_0$  against the alternative hypothesis  $\mu_X > \mu_Y$ , we are performing a one-sided test: We use the same test statistic  $Z$ , but we reject values of  $Z$  which correspond to large values of  $\mu_X - \mu_Y$ , so large positive values of  $Z$ . Large negative values of  $Z$ , while they are unlikely for the null hypothesis, are even more unlikely for the alternative. The cutoff for testing at the 0.05 level is  $z_{0.95} = 1.64$ . Thus, we do not reject the null hypothesis.

### 11.3 Z test for the difference between proportions

Proportions are analysed in exactly the same way as population means, where the population consists only of the numbers 0 and 1 in some unknown proportion. We sample two populations,  $x$  and  $y$ , with  $n_x$  samples and  $n_y$  samples respectively, and observe  $k_x$  and  $k_y$  “successes” respectively. Under the null hypothesis, the population SD is  $\sqrt{p(1-p)}$ , where  $p$  is the common proportion of 1’s (successes) in the population. We substitute the estimate from the sample  $\hat{p} = (k_x + k_y)/(n_x + n_y)$ .

Consider the results of a study that was carried out in Rakai, Uganda to test the theory that circumcision would reduce infection rates for HIV. While the procedure seemed to succeed in its primary goal — reducing infection rates of the men who were circumcised — there was some evidence that it actually increased the likelihood of the men’s partners becoming infected. The results (reported in The International Herald Tribune 6 March, 2007) showed that among 70 men with HIV who were circumcised, 11 of their partners became infected in the month following surgery; among 54 controls who were not circumcised, only 4 of the partners became infected in the first month. Writing subscripts c and u for “circumcised” and “uncircumcised” respectively, we have then the estimated proportion infected  $p_c = 11/70 = 0.157$ , and  $p_u = 4/54 = 0.074$ . Could the difference be simply due to chance? We perform a Z test at the 0.05 significance level.

The joint estimate of the proportion is  $\hat{p} = 15/124 = 0.121$ , giving us a sample SD of  $\hat{\sigma} = 0.326$ . The standard error for the difference is then

$$SE_{P_u - P_c} = \hat{\sigma} \sqrt{\frac{1}{n_u} + \frac{1}{n_c}} = 0.326 * \sqrt{\frac{1}{54} + \frac{1}{70}} = 0.059.$$

The z statistic is then

$$Z = \frac{p_u - p_c}{SE} = \frac{-0.083}{0.059} = -1.41.$$

The cutoff for rejecting Z at the 0.05 level is 1.96. Since the observed Z is smaller than this, we do not reject the null hypothesis, that the infection rates are in fact equal. The difference in infection rates is *not statistically significant*, as we cannot be confident that the difference is not simply due to chance.

## 11.4 t confidence interval for the difference between population means

Consider the following study [SCT<sup>+</sup>90], discussed in [RS02, Chapter 2]: Researchers measured the volume of the left hippocampus of the brains of 30 men, of whom 15 were schizophrenic. The goal is to determine whether there is a difference in the size of this brain region between schizophrenics and unaffected individuals. The data are given in Table 11.2. The average size among the unaffected subjects (in cm<sup>3</sup>) is 1.76, while the mean for the schizophrenic subjects is 1.56. The sample SDs are 0.24 and 0.3 respectively. What can we infer about the populations that these individuals were sampled from? Do schizophrenics have smaller hippocampal volume, on average?

We make the modeling assumption that these individuals are a random sample from the general population (of healthy and schizophrenic men, respectively. More about this in section 11.5.2.) We also assume that the underlying variance of the two populations is the same, but unknown: The difference between the two groups (potentially) is in the population means  $\mu_x$  (healthy) and  $\mu_y$  (schizophrenic), and we want a confidence interval for the difference.

Since we don't know the population SD in advance, and since the number of samples is small, we use the T distribution for our confidence intervals instead of the normal. (Since we don't know that the population is normally distributed, we are relying on the normal approximation, which may be questionable for averages small samples. For more about the validity of this assumption, see section 11.5.3.) As always, the symmetric 95% confidence interval is of the form

$$\text{Estimate} \pm t \times SE,$$

where  $t$  is the number such that 95% of the probability in the appropriate T distribution is between  $-t$  and  $t$  (that is, the number in the  $P = 0.05$  column of your table.) We need to know

- (1). How many degrees of freedom?
- (2). What is the SE?

The first is easy: We add the degrees of freedom, to get  $n_x + n_y - 2$  — in this case, 28.

The second is sort of easy: Like for the Z test, when  $\sigma_x = \sigma_y$  (that's

what we're assuming), we get

$$SE = \sigma \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}.$$

The only problem is that we don't know what  $\sigma$  is. We have our sample SDs  $s_x$  and  $s_y$ , each of which should be approximately  $\sigma$ . The bigger the sample, the better the approximation should be. This leads us to the **pooled sample variance**  $s_p^2$ , which simply averages these estimates, counting the bigger sample more heavily:

<b>pooled sample variance</b> : $s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}$ .
$SE = s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$

Plugging in the data, we get  $s_p = 0.27$ , so that the SE becomes 0.099. The table gives us  $t = 2.05$  (with 28 d.f.), so that the 95% confidence interval for  $\mu_x - \mu_y$  becomes

$$0.20 \pm 2.05 \times 0.099 = (-0.003, 0.403).$$

Table 11.2: Data from the Suddath [RS90] schizophrenia experiment. Hippocampus volumes in  $cm^3$ .

Unaffected :	1.94, 1.44, 1.56, 1.58, 2.06, 1.66, 1.75, 1.77, 1.78, 1.92, 1.25, 1.93, 2.04, 1.62, 2.08;
Schizophrenic :	1.27, 1.63, 1.47, 1.39, 1.93, 1.26, 1.71, 1.67, 1.28, 1.85, 1.02, 1.34, 2.02, 1.59, 1.97.

## 11.5 Two-sample test and paired-sample test.

### 11.5.1 Schizophrenia study: Two-sample t test

We observe that the confidence interval computed in section 11.4 includes 0, meaning that we are not 95% confident that the difference is not 0. If this is the question of primary interest, we can formulate this as a hypothesis test.

Is the difference in hippocampal volume between the two groups statistically significant? Can the observed difference be due to chance? We perform a t test at significance level 0.05. Our null hypothesis is that  $\mu_x = \mu_y$ , and our two-tailed alternative is  $\mu_x \neq \mu_y$ . The standard error is computed exactly as before, to be 0.099. The T test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{SE} = \frac{0.20}{0.099} = 2.02.$$

We then observe that this is not above the critical value 2.05, so we RETAIN the null hypothesis, and say that the difference is **not statistically significant**.

If we had decided in advance that we were only interested in whether  $\mu_x > \mu_y$  — the one-tailed alternative — we would use the same test statistic  $T = 2.02$ , but now we draw our critical value from the  $P = 0.10$  column, which gives us 1.70. In this case, we would reject the null hypothesis. On the other hand, if we had decided in advance that our alternative was  $\mu_x < \mu_y$ , we would have a critical value  $-1.70$ , with rejection region anything below that, so of course we would retain the null hypothesis.

### 11.5.2 The paired-sample test

It may seem disappointing that we can't do more with our small samples. The measurements in the healthy group certainly *seem* bigger than those of the schizophrenic group. The problem is, there's so much noise, in the form of general overall variability among the individuals, that we can't be sure if the difference between the groups is just part of that natural variation.

It might be nice if we could get rid of some of that noise. For instance, suppose the variation between people was in three parts — call them  $A$ ,  $B$ , and  $S$ , so your hippocampal volume is  $A + B + S$ .  $S$  is the effect of having schizophrenia, and  $A$  is random stuff we have no control over.  $B$  is another individual effect, but one that we suppose is better defined — for instance, the effect of genetic inheritance. Suppose we could pair schizophrenic and non-schizophrenic people with the same  $B$  score up, and then look at the difference between individuals within a pair. Then  $B$  cancels out, and  $S$  becomes more prominent. This is the idea of the **matched case-control study**.

In fact, that's just what was done in this study. The real data are given in Table 12.5. The 30 subjects were, in fact, 30 pairs of monozygotic twins, of whom one was schizophrenic and the other not. The paired-sample T test is exactly the one that we described in section 10.2. The mean of the

differences is 0.20, and the sample SD of the fifteen differences is 0.238. The SE for these differences is then  $0.238/\sqrt{15} = 0.0615$ . In other words, while the average differences between 15 independent pairs of schizophrenic and healthy subject hippocampus volumes would vary by about 0.099, the differences in our sample vary by only about 0.0615 — so, about 40% less — because some of the variability has been excluded by matching the individuals in a pair.

We compute then  $T = 0.20/0.0615 = 3.25$ . Since the critical value in 2.15, for T with 14 degrees of freedom at the 0.05 level, we clearly reject the null hypothesis, and conclude that there is a significant difference between the schizophrenic and unaffected brains. (Of course, we do have to ask whether results about twins generalise to the rest of the population.)

Table 11.3: Data from the Suddath [RS90] schizophrenia experiment. Hippocampus volumes in  $cm^3$ .

Unaffected	Schizophrenic	Difference
1.94	1.27	0.67
1.44	1.63	-0.19
1.56	1.47	0.09
1.58	1.39	0.19
2.06	1.93	0.13
1.66	1.26	0.40
1.75	1.71	0.04
1.77	1.67	0.10
1.78	1.28	0.50
1.92	1.85	0.07
1.25	1.02	0.23
1.93	1.34	0.59
2.04	2.02	0.02
1.62	1.59	0.03
2.08	1.97	0.11

**General rule:** Suppose you wish to do a Z or T test for the difference between the means of two normally distributed populations. If the data are naturally paired up, so that the two observations in a pair are positively correlated (see chapter 15 for precise definitions), then it makes sense to compute the differences first, and then perform a one-sample test on the differences. If not, then we perform the two-sample Z or T test, depending on the circumstances.

You might imagine that you are sampling at random from a box full of

cards, each of which has an X and a Y side, with numbers on each, and you are trying to determine from the sample whether X and Y have the same average. You could write down your sample of X's, then turn over all the cards and write down your sample of Y's, and compare the means. If the X and Y numbers tend to vary together, though, it makes more sense to look at the differences  $X - Y$  over the cards, rather than throw away the information about which X goes with which Y. If the X's and Y's are not actually related to each other then it shouldn't matter.

### 11.5.3 Is the CLT justified?

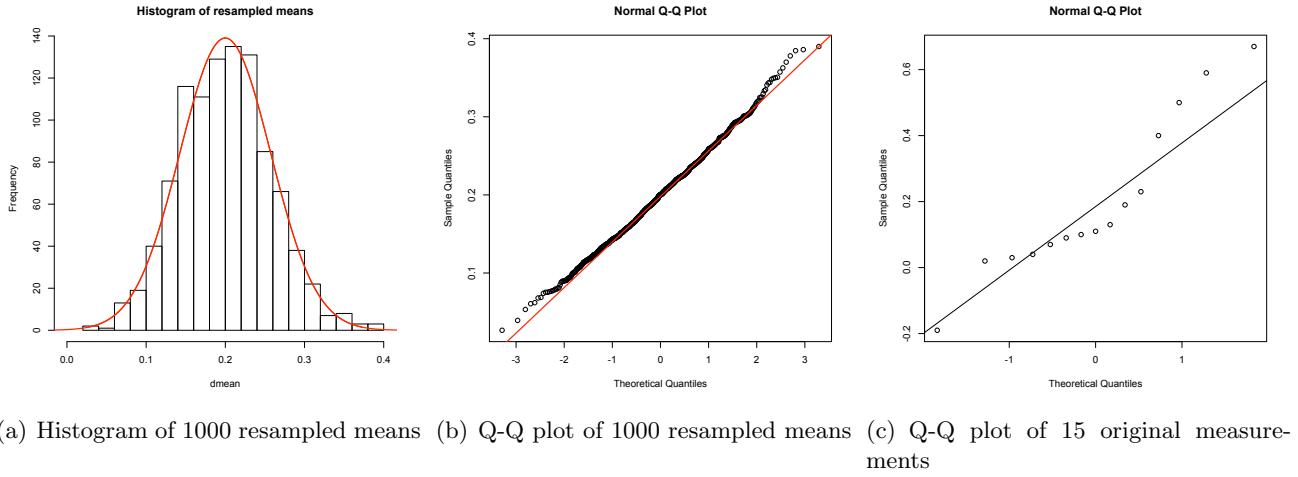
In section 11.5.2 we supposed that the average difference between the unaffected and schizophrenic hippocampus volumes would have a nearly normal distribution. The CLT tells us that that the average of a very large number of such differences, picked at random from the same distribution, would have an approximately normal distribution. But is that true for just 15? We would like to test this supposition.

One way to do this is with a random experiment. We sample 15 volume differences at random from the whole population of twins, and average them. Repeat 1000 times, and look at the distribution of averages we find. Are these approximately normal?

*But wait! We don't have access to any larger population of twins; and if we did, we would have included their measurements in our study. The trick (which is widely used in modern statistics, but is not part of this course), called "the bootstrap", is instead to resample from the data we already have, picking 15 samples with replacement from the 15 we already have — so some will be counted several times, and some not at all. It sounds like cheating, but it can be shown mathematically that it works.*

A histogram of the results is shown in Figure 11.2(a), together with the appropriate normal curve. The fit looks pretty good, which should reassure us of the appropriateness of the test we have applied. Another way of seeing that this distribution is close to normal is Figure 11.2(b), which shows a so-called Q-Q plot. (The Q-Q plot is not examinable as such. In principle, it is the basis for the Kolmogorov-Smirnov test, which we describe in section 13.1. This will give us a quantitative answer to the question: Is this distribution close to the normal distribution?) The idea is very simple: We have 1000 numbers that we think might have been sampled from a normal distribution. We look at the normal distribution these might have been sampled from —

the one with the same mean and variance as this sample — and take 1000 numbers evenly spaced from the normal distribution, and plot them against each other. If the sample really came from the normal distribution, then the two should be about equal, so the points will all lie on the main diagonal. Figure 11.2(c) shows a Q-Q plot for the original 15 samples, which clearly do not fit the normal distribution very well.



(a) Histogram of 1000 resampled means (b) Q-Q plot of 1000 resampled means (c) Q-Q plot of 15 original measurements

Figure 11.2: Comparisons of normal distribution to means of resampled schizophrenia data and original schizophrenia data.

## 11.6 Hypothesis tests for experiments

### 11.6.1 Quantitative experiments

276 women were enrolled in a study to evaluate a weight-loss intervention program [SKK91]. They were allocated at random to one of two different groups: 171 women in the intervention group received nutrition counseling and behaviour modification treatment to help them reduce the fat in their diets. The 105 women in the control group were urged to maintain their current diet.

After 6 months, the intervention group had lost 3.2kg on average, with an SD of 3.5 kg, while the control group had lost 0.4kg on average, with an SD of 2.8 kg. Is the difference in weight loss between the two groups statistically significant?

Let us first apply a Z test at the 0.05 significance level, without thinking too deeply about what it means. (Because of the normal approximation, it doesn't really matter what the underlying distribution of weight loss is.) We compute first the pooled sample variance:

$$s_p^2 = \frac{170 \cdot 3.5^2 + 104 \cdot 2.8^2}{274} = 10.6,$$

so  $s_p = 3.25\text{kg}$ . The standard error is  $s_p \sqrt{1/n + 1/m} = 0.40\text{kg}$ . Our test statistic is then

$$Z = \frac{\bar{x} - \bar{y}}{SE} = \frac{3.1}{0.4} = 7.7.$$

This exceeds the rejection threshold of 1.96 by a large margin, so we conclude that the difference in weight loss between the groups is statistically significant.

But are we justified in using this hypothesis test? What is the random sampling procedure that defines the null hypothesis? What is the “just by chance” that could have happened, and that we wish to rule out? These 276 women were not randomly selected from any population — at least, not according to any well-defined procedure. The randomness is in the assignment of women to the two groups: We need to show that the difference between the two groups is not merely a chance result of the women that we happened to pick for the intervention group, which might have turned out differently if we happened to pick differently.

In fact, this model is a lot like the model of section 11.5. Imagine a box containing 276 cards, one for each woman in the study. Side A of the card says how much weight the woman would lose if she were in the intervention group; side B says how much weight she would lose if she were in the control group. The null hypothesis states, then, that the average of the A's is the same as the average of the B's. The procedure of section 11.5 says that we compute all the values A-B from our sample, and test whether these could have come from a distribution with mean 0. The problem is that we never get to see A and B from the same card.

Instead, we have followed the procedure of section 11.5.1, in which we take a sample of A's and a sample of B's, and test for whether they could have come from distributions with the same mean. But there are some important problems that we did not address there:

- We sample the A's (the intervention group) without replacement. Furthermore, this is a large fraction of the total population (in the box). We know from section 10.3.1 that this makes the SE **smaller**.

- The sample of B's is not an independent sample; it's just the complement of the sample of A's. A bit of thought makes clear that this tends to make the SE of the difference **larger**.

This turns out to be one of those cases where two wrongs really do make a right. These two errors work in opposite directions, and pretty much cancel each other out. Consequently, in analysing experiments we ignore these complications and proceed with the Z- or t-test as in section 11.5.1, as though they were independent samples.

### 11.6.2 Qualitative experiments

A famous experiment in the psychology of choice was carried out by A. Tversky and D. Kahnemann [TK81], to address the following question: Do people make economic decisions by a rational calculus, where they measure the perceived benefit against the cost, and then choose the course of action with the highest return? Or do they apply more convoluted decision procedures? They decided to try to find out whether people would give the same answer to the following questions:

**Question A:** Imagine that you have decided to see a play where admission is \$10 per ticket. As you enter the theatre you discover that you have lost a \$10 bill. Would you still pay \$10 for a ticket for the play?

**Question B:** Imagine that you have decided to see a play and paid the admission price of \$10 per ticket. As you enter the theatre you discover that you have lost the ticket. The seat was not marked and the ticket cannot be recovered. Would you pay \$10 for another ticket?

From a rational economic perspective, the two situations are exactly identical, from the point of view of the subject: She is at the theatre, the play that she wants to see is about to start, but she doesn't have a ticket and would have to buy one. But maybe people still see these two situations differently. The problem is, you can't just show people both questions and ask them if they would answer the same to both questions. Instead, they posed Question A to about half the subjects (183 people) and Question B to the other half (200 people). The results are given in Table 11.4.

It certainly *appears* that people are more likely to answer yes to A than to B (88% vs. 46%), but could this difference be merely due to chance? As usual, we need to ask, what is the chance model? It is not about the sample of 383 people that we are studying. They are not a probability sample from

Table 11.4

	Yes	No
Question A	161	22
Question B	92	54

any population, and we have no idea how representative they may or may not be of the larger category of *Homo sapiens*. The real question is, among these 383 people, how likely is it that we would have found a different result had we by chance selected a different group of 200 people to pose question B to. We want to do a significance test at the 0.01 level.

The model is then: 383 cards in a box. On one side is that person's answer to Question A, on the other side the same person's answer to Question B (coded as 1=yes, 0=no). The null hypothesis is that the average on the A side is the same as the average on the B side (which includes the more specific hypothesis that the A's and the B's are identical).

We pick 183 cards at random, and add up their side A's, coming to 161; from the other 200 we add up the side B's, coming to 92. Our procedure is then:

- (1). The average of the sampled side A's is  $\bar{X}_A = 0.88$ , while the average of the sampled side B's is  $\bar{X}_B = 0.46$ .
- (2). The standard deviation of the A sides is estimated at  $\sigma_A = \sqrt{p(1-p)} = 0.32$ , while the standard deviation of the B sides is estimated at  $\sigma_B = \sqrt{p(1-p)} = 0.50$ .
- (3). The standard error for the difference is estimated at

$$SE_{A-B} = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}} = \sqrt{\frac{0.32^2}{183} + \frac{0.5^2}{200}} = 0.043.$$

- (4).  $Z = (\bar{X}_A - \bar{X}_B)/SE_{A-B} = 9.77$ . The cutoff for a two-sided test at the 0.01 level is  $z_{0.995} = 2.58$ , so we clearly do reject the null hypothesis.

The conclusion is that the difference in answers between the two questions was not due to the random sampling. Again, this tells us nothing directly about the larger population from which these 383 individuals were sampled.



## Lecture 12

# Non-Parametric Tests, Part I

### 12.1 Introduction: Why do we need distribution-free tests?

One of the most common problems we need to deal with in statistics is to compare two population means: We have samples from two different populations, and we want to determine whether the populations they were drawn from could have had the same mean, or we want to compute a confidence interval for the difference between the means. The methods described earlier in the course began with the assumption that the populations under consideration were normally distributed, and only the means (and perhaps the variances) were unknown. But most data that we consider do not come from a normal distribution. What do we do then?

In section 7.4 we discussed how we can use a mathematical result — the Central Limit Theorem — to justify treating data as though they had come from a normal distribution, as long as enough independent random samples are being averaged. But in some cases we don't have enough samples to invoke the Central Limit Theorem. In other cases (such as that of section 11.5.1) the experimental design seems to lack the randomisation that would justify the normal approximation.

In this lecture and the next we describe alternative approaches to the standard hypothesis tests described in previous lectures, which are independent of any assumption about the underlying distribution of the data. In the following lectures we describe new problems — partitioning variance into different sources, and describing the strength of relationship between different variables — and in sections 14.6 and 16.3.2 we will present non-parametric versions of these.

The advantage of these “non-parametric tests” lie in their robustness: The significance level is known, independent of assumptions about the distribution from which the data were drawn — a valuable guarantee, since we can never really confirm these assumptions with certainty. In addition, the non-parametric approach can be more logically compelling for some common experimental designs, something we will discuss further in section 12.4.4.

Of course, there is always a tradeoff. The reliability of the non-parametric approach comes at the expense of power: The non-parametric test is always less powerful than the corresponding parametric test. (Or, to put it the other way, if we know what the underlying distribution is, we can use that knowledge to construct a more powerful test at the same level as the generic test that works for any distribution.)

## 12.2 First example: Learning to Walk

### 12.2.1 A first attempt

We recall the study of infant walking that we described way back in section 1.1. Six infants were given exercises to maintain their walking reflex, and six control infants were observed without any special exercises. The ages (in months) at which the infants were first able to walk independently are recapitulated in Table 12.1.

Treatment	9.00	9.50	9.75	10.00	13.00	9.50
Control	11.50	12.00	9.00	11.50	13.25	13.00

Table 12.1: Age (in months) at which infants were first able to walk independently. Data from [ZK72].

As we said then, the Treatment numbers seem generally smaller than the Control numbers, but not entirely, and the number of observations is small. Could we merely be observing sampling variation, where we happened to get six (five, actually) early walkers in the Treatment group, and late walkers in the Control group.

Following the approach of Lecture 11, we might perform a two-sample T test for equality of means. We test the null hypothesis  $\mu_{TREAT} = \mu_{CON}$  against the one-tailed alternative  $\mu_{TREAT} < \mu_{CON}$ , at the 0.05 level. To find the critical value, we look in the column for  $P = 0.10$ , with  $6 + 6 - 2$  d.f., obtaining 1.81. The critical region is then  $\{T < -1.81\}$ . The relevant

summary statistics are given in Table 12.2. We compute the pooled sample variance

$$s_p = \sqrt{\frac{(6-1)1.45^2 + (6-1)1.52^2}{6+6-2}} = 1.48,$$

so the standard error is

$$SE = s_p \sqrt{\frac{1}{6} + \frac{1}{6}} = 0.85.$$

We have then the T statistic

$$T = \frac{\bar{X} - \bar{Y}}{SE} = \frac{-1.6}{0.85} = -1.85.$$

So we reject the null hypothesis, and say that the difference between the two groups is statistically significant.

	Mean	SD
Treatment	10.1	1.45
Control	11.7	1.52

Table 12.2: Summary statistics from Table 12.1

### 12.2.2 What could go wrong with the T test?

We may wonder about the validity of this test, though, particularly as the observed T was just barely inside the critical region. After all, the T test depends on the assumption that when  $X_1, \dots, X_6$  and  $Y_1, \dots, Y_6$  are independent samples from a **normal distribution** with unknown mean and variance, then the observed T statistic will be below  $-1.81$  just 5% of the time. But the data we have, sparse though they are, don't look like they come from a normal distribution. They look more like they come from a bimodal distribution, like the one sketched in Figure 12.1.

So, there might be early walkers and late walkers, and we just happened to get mostly early walkers for the Treatment group, and late walkers for the Control group. How much does this matter? We present one way of seeing this in section 12.2.3. For the time being, we simply note that there is a potential problem, since the whole idea of this statistical approach was to develop some certainty about the level of uncertainty.

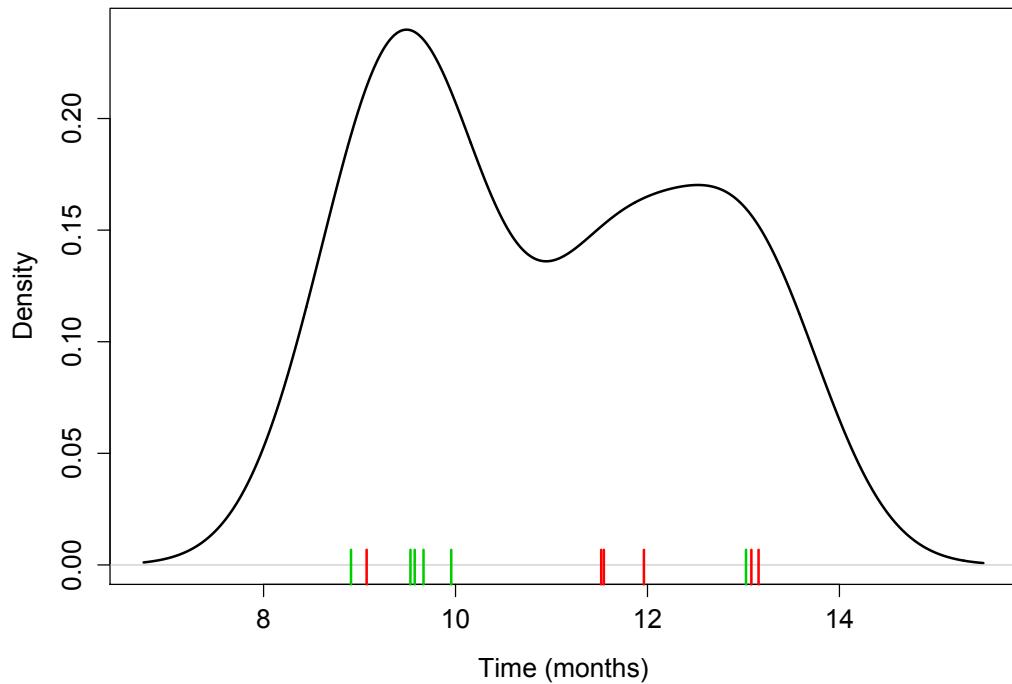


Figure 12.1: Sketch of what the distribution of walking times from which the data of Table 12.1 might have been drawn from, if they all came from the same distribution. The actual measurements are shown as a rug plot along the bottom — green for Treatment, red for Control. The marks have been adjusted slightly to avoid exact overlaps.

We would like to have an alternative procedure for testing hypotheses about equality of distributions, which will give correct significance levels without depending on (possibly false) assumptions about the shape of the distributions. Of course, you never get something for nothing. In return for having procedures that are more generally applicable, and give the correct significance level when the null hypothesis is true, we will lose some power: The probability of rejecting a false null hypothesis will be generally lower, so that we will need larger samples to attain the same level of confidence. Still, true confidence is better than deceptive confidence! We call these alternatives “distribution-free” or “non-parametric” tests. Some possibilities are described in section 12.3.

### 12.2.3 How much does the non-normality matter?

*This section is optional*

Suppose the null hypothesis were true, that the samples really came all from the same distribution, but that distribution were not normal, but distributed like the density in Figure 12.1. What would the distribution of the T statistics then look like? We can find this out by simulation: We pick 2 groups of six at random from this distribution, and call one of them “Treatment”, and the other “Control”. We compute T from these 12 observations, exactly as we did on the real data, and then repeat this 10,000 times.

The histogram of these 10,000 simulated T values is shown in Figure 12.2. Notice that this is similar to the superimposed Student T density, but not exactly the same. In fact, in the crucial tails there are substantial differences. Table 12.3 compares the critical values at different significance levels for the theoretical T distribution and the simulated (non-normal) T distribution. Thus, we see in Table 12.3(b) that if we do a one-tailed test at the 0.05 significance level by rejecting  $T < -1.81$ , instead of making the (correct) 10% type I errors, we will make 10.4% type I errors. Similarly, if we reject  $|T| > 3.17$ , to make what we think is a two-tailed test at level 0.01, we will in fact be making 1.1% type I errors when the null hypothesis is true. Thus, we would be overstating our confidence.

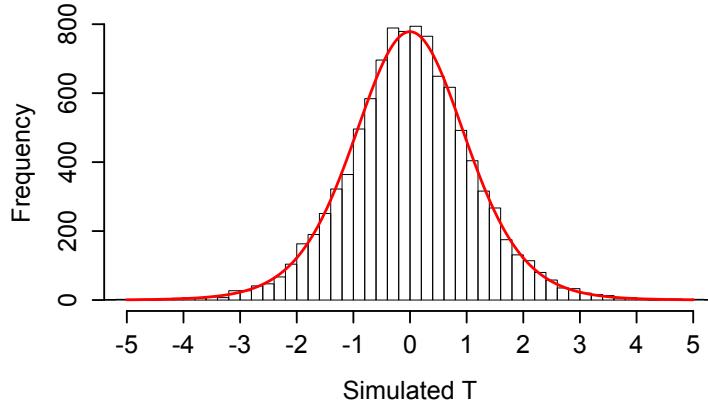


Figure 12.2: Histogram of 10,000 simulated T values, computed from pairs of samples of six values each from the distribution sketched in Figure 12.1. The Student T density with 10 degrees of freedom is superimposed in red.

Table 12.3: Comparison of the tail probabilities of the T distribution simulated from non-normal samples, with the standard T distribution.

(a) Real vs. standard critical values			(b) Real vs. standard tail probabilities		
P	Standard	Simulated	T	Standard	Simulated
0.10	1.81	1.83	1.81	0.100	0.104
0.05	2.23	2.25	2	0.073	0.076
0.01	3.17	3.22	2.5	0.031	0.035
0.001	4.59	5.14	3.17	0.010	0.011
			4	0.0025	0.0033

Why are these tests called **non-parametric**? The more basic hypothesis tests that you have already learned are **parametric**, because they start from the assumption that both populations have distributions that fall into the same general class — generally normal — and the specific member of that class is determined by a single number, or a few numbers — the *parameters*. We are testing to decide only whether the parameters are the same. In the new tests, we drop this assumption, allowing *a priori* that the two population distributions could be anything at all.

## 12.3 Tests for independent samples

### 12.3.1 Median test

Suppose we have samples  $x_1, \dots, x_{n_x}$  and  $y_1, \dots, y_{n_y}$  from two distinct distributions whose medians are  $m_x$  and  $m_y$ . We wish to test the null hypothesis

$$H_0 : m_x = m_y$$

against a two-tailed alternative

$$H_{alt} : m_x \neq m_y,$$

or a one-tailed alternative

$$H_{alt} : m_x < m_y, \quad \text{or } H_{alt} : m_x > m_y.$$

The idea of this test is straightforward: Let  $M$  be the median of the combined sample  $\{x_1, \dots, x_{n_x}, y_1, \dots, y_{n_y}\}$ . If the medians are the same, then the  $x$ 's and the  $y$ 's should have an equal chance of being above  $M$ . Let  $P_x$  be the proportion of  $x$ 's that are above  $M$ , and  $P_y$  the proportion of  $y$ 's that are above  $M$ . It turns out that we can treat these as though they were the proportions of successes in  $n_x$  and  $n_y$  trials respectively. Analysing these results is not entirely straightforward; we get a reasonable approximation by using the Z test for differences between proportions, as in section 11.3.

Consider the case of the infant walking study, described in section 1.1. The 12 measurements are

9.0, 9.0, 9.5, 9.5, 9.75, 10.0, 11.5, 11.5, 12.0, 13.0, 13.0, 13.25,

where the **CONTROL** results have been coloured **RED**, and the **TREATMENT** results have been coloured **GREEN**. The median is 10.75, and we see that there are 5 control results above the median, and one treatment.

#### Calculating the p-value: Exact method

Imagine that we have  $n_x$  red balls and  $n_y$  green balls in a box. We pick half of them at random — these are the above median outcomes — and get  $k_x$  red and  $k_y$  green. We expected to get about  $n_x/2$  red and  $n_y/2$  green. What is the probability of such an extreme result? This is reasonably straightforward to compute, though slightly beyond what we're doing in this course. We'll just go through the calculation for this one special case.

We pick 6 balls from 12, where 6 were red and 6 green. We want  $P(\text{at least 5 red})$ . Since the null hypothesis says that all picks are equally likely, this is simply the fraction of ways that we could make our picks which happen to have 5 or 6 red. That is,

$$P(\text{at least 5 R}) = \frac{\# \text{ ways to pick 5 R, 1G} + \# \text{ ways to pick 6 R, 0G}}{\text{total } \# \text{ ways to pick 6 balls from 12}}.$$

The number of ways to pick 6 balls from 12 is what we call  ${}^{12}C_6 = 924$ . The number of ways to pick 6 red and 0 green is just 1: We have to take all the reds, we have no choice. The only slightly tricky one is  $\#$  ways to pick 5 red and 1 green. A little thought shows that we have  ${}^6C_5 = 6$  ways of choosing the red balls, and  ${}^6C_1 = 6$  ways of choosing the green one, so 36 ways in all. Thus, the p-value comes out to  $37/954 = 0.039$ , so we still reject the null hypothesis at the 0.05 level. (Of course, the p-value for a two-tailed test is twice this, or 0.078.

### Calculating the p-value: Approximate method

Slightly easier is to use the normal approximation to compute the p-value. This is the method described in your formula booklets. But this needs to be done with some care.

The idea is the following: We have observed a certain number  $n_+$  of above-median outcomes. (This will be about  $\frac{1}{2}$  the total number of samples  $n = n_x + n_y$ , but slightly adjusted if the number is odd, or if there are ties.) We have observed proportions  $p_x = k_x/n_x$  and  $p_y = k_y/n_y$  above-median samples from the two groups, respectively. We apply a Z test (as in sections 11.3 and 11.6.2) to test whether these proportions could be “really” the same (as they should be under the null hypothesis).<sup>1</sup>

We compute the standard error as

$$SE = \sqrt{\hat{p}(1 - \hat{p})} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}},$$

---

<sup>1</sup>If you’re thinking carefully, you should object at this point, that we can’t apply the Z test for difference between proportions, because that requires that the two samples be independent. That’s true. The fact that they’re dependent raises the standard error; but the fact that we’re sampling without replacement lowers the standard error. The two effects more or less cancel out, just as we described in section 11.6. Another way of computing this result is to use the  $\chi^2$  test for independence, writing this as a  $2 \times 2$  contingency table. In principle this gives the same result, but it’s hard to apply the continuity correction there. In a case like this one, the expected cell occupancies are smaller than we allow.

where  $\hat{p} = n_+/n$  is the proportion of samples above the median, which is  $\frac{1}{2}$  or a bit below. Then we compute the test statistic

$$Z = \frac{p_x - p_y}{SE},$$

and find the p-value by looking this up on the standard normal table.

Applying this to our infant walking example, we see obtain  $\hat{p} = \frac{1}{2}$ , and

$$SE = \frac{1}{2} \sqrt{\frac{1}{6} + \frac{1}{6}} = 0.289.$$

Then we get  $Z = -0.667/0.289 = 2.3$ . The normal table gives us a p-value of 0.01 for the one-tailed test. This is a long way off our exact computation of 0.04. What went wrong?

The catch is that we are approximating a discrete distribution with a continuous one (the normal), and that can mean substantial errors when the numbers involved are small, as they are liable to be when we are interested in applying the median test. We need a continuity correction (as discussed in section 6.8.1). We are asking for the probability of having at least 5 out of 6 from the Control group. But if we are thinking of this as a continuous measurement, we have to represent it as at least 4.5 out of 6. (A picture of the binomial probability with the normal approximation is given in Figure 12.3.) Similarly, the extreme fraction of treatment samples in the above-median sample must be seen to start at  $1.5/6 = 0.25$ , rather than at  $1/6$ . Thus, we have the test statistic

$$Z = \frac{0.25 - 0.75}{0.289} = -1.73.$$

If we look up 1.73 on the normal table, we get the value 0.9582, meaning that the one-tailed probability below  $-1.73$  is  $1 - 0.9582 = 0.0418$ , which is very close to the exact value computed above. Thus, we see that this normal approximation can work very well, if we remember to use the continuity correction.

To compute the observed significance level (p-value) for the median test, we use the Z test for differences between proportions, applied to the proportions  $p_x$  and  $p_y$  of the two samples that are above the median,

$Z = (p_x - p_y)/SE$ , with  $SE = \sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_x} + \frac{1}{n_y} \right)}$ . If  $p_x$  is the larger proportion, we adjust it to

$$p_x = \frac{k_x - 0.5}{n_+} \quad \text{and} \quad p_y = \frac{k_y + 0.5}{n_+}.$$

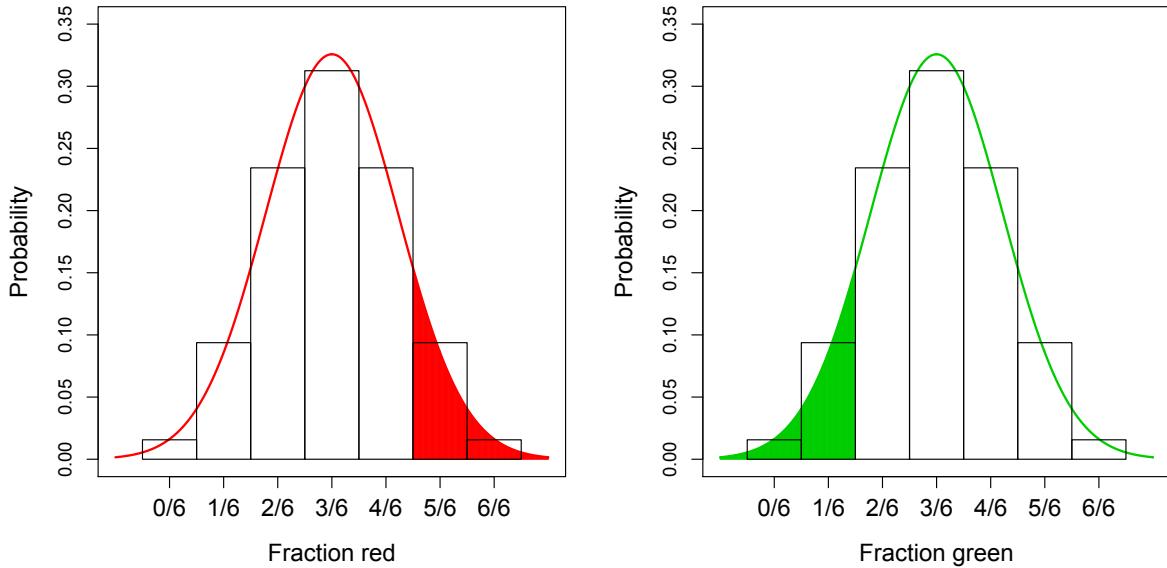


Figure 12.3: The exact probabilities from the binomial distribution for the extreme results of number of red (control) and green (treatment) in the infant walking experiment. The corresponding normal approximations are shaded. Note that the upper tail starts at  $4.5/6 = 0.75$ , not at  $5/6$ ; and the lower tail starts at  $1.5/6 = 0.25$ , rather than at  $1/6$ .

There are many defects of the median test. One of them is that the results are discrete — there are at most  $n/2 + 1$  different possible outcomes to the test — while the analysis with  $Z$  is implicitly discrete. This is one of the many reasons why the median test, while it is sometimes seen, is not recommended. (For more about this, see [FG00].) The rank-sum test is almost always preferred.

Note that this method requires that the observations be all distinct. There is a version of the median test that can be used when there are ties among the observations, but we do not discuss it in this course.

### 12.3.2 Rank-Sum test

The median test is obviously less powerful than it could be, because it considers only how many of each group are above or below the median, but

not how far above or below. In the example of section 12.2, while 5 of the 6 treatment samples below the median, the one that is above the median is near the top of the whole sample; and the one control sample that is below the median is in fact near the bottom. It seems clear that we should want to take this extra information into account. The idea of the rank-sum test (also called the Mann-Whitney test) is that we consider not just yes/no, above/below the median, but the exact relative ranking.

Continuing with this example, we list all 12 measurements in order, and replace them by their ranks:

measurements	9.0	9.0	9.5	9.5	9.75	10.0	11.5	11.5	12.0	13.0	13.0	13.25
ranks	1	2	3	4	5	6	7	8	9	10	11	12
modified ranks	1.5	1.5	3.5	4.5	5	6	7.5	7.5	9	10.5	10.5	12

When measurements are tied, we average the ranks (we show this in the column labelled “modified ranks”.) We wish to test the null hypothesis  $H_0$  : control and treatment came from the same distribution; against the alternative hypothesis that the controls are generally larger.

We compute a test statistic  $R$ , which is just the sum of the ranks in the smaller sample. (In this case, the two samples have the same size, so we can take either one. We will take the treatment sample.) The idea is that these should, if  $H_0$  is true, be like a random sample from the numbers  $1, \dots, n_x + n_y$ . If  $R$  is too big or too small we take this as evidence to reject  $H_0$ . In the one-tailed case, we reject  $R$  for being too small (if the alternative hypothesis is that the corresponding group has *smaller* values; or for being too large (if the alternative hypothesis is that the corresponding group has *larger* values.

In this case, the alternative hypothesis is that the group under consideration, the treatment group has smaller values, so the rejection region consists of  $R$  below a certain threshold. It only remains to find the appropriate threshold. These are given on the Mann-Whitney table (Table 5 in the formula booklet). The layout of this table is somewhat complicated. The table lists critical values corresponding only to  $P = 0.05$  and  $P = 0.10$ . We look in the row corresponding to the size of the smaller sample, and column corresponding to the larger. For a two-tailed test we look in the (sub-) row corresponding to the desired significance level; for a one-tailed test we double the p-value.

The sum of the ranks for the treatment group in our example is  $R = 30$ . Since we are performing a one-tailed test with the alternative being that the treatment values are **smaller**, our rejection region will be of the form

$R \leq$  some critical value. We find the critical value on Table 12.4. The values corresponding to two samples of size 6 have been highlighted. For a one-tailed test at the 0.05 level we take the upper values 28, 50. Hence, we would reject  $R \leq 28$ . Since  $R = 30$ , we retain the null hypothesis in this test.

If we were performing a two-tailed test instead, we would reject  $R \leq 26$  and  $R \geq 52$ .

The table you are given goes up only as far as the larger sample size equal to 10. For larger samples, we use a normal approximation:

$$\begin{aligned} z &= \frac{R - \mu}{\sigma}, \\ \mu &= \frac{1}{2}n_x(n_x + n_y + 1), \\ \sigma &= \sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}}. \end{aligned}$$

As usual, we compare this  $z$  to the probabilities on the normal table. Thus, for instance, for a two-tailed test at the 0.05 level, we reject the null hypothesis if  $|z| > 1.96$ .

		larger sample size, $n_2$						
		4	5	6	7	8	9	10
smaller sample size $n_1$	4	12,24 11,25	13,27 12,28	14,30 12,32	15,33 13,35	16,36 14,38	17,39 15,41	18,42 16,44
	5		19,36 18,37	20,40 19,41	22,43 20,45	23,47 21,49	25,50 22,53	26,54 24,56
6	6			28,50 26,52	30,54 28,56	32,58 29,61	33,63 31,65	35,67 33,69
	7				39,66 37,68	41,71 39,73	43,76 41,78	46,80 43,83
8	8					52,84 49,87	54,90 51,93	57,95 54,98
	9						66,105 63,108	69,111 66,114
10	10							83,127 79,131

Table 12.4: Critical values for Mann-Whitney rank-sum test.

## 12.4 Tests for paired data

As with the t test, when the data fall naturally into matched pairs, we can improve the power of the test by taking this into account. We are given data in pairs  $(x_1, y_1), \dots, (x_n, y_n)$ , and we wish to test the null hypothesis  $H_0$ : x and y come from the same distribution. In fact, the null hypothesis may be thought of as being even broader than that. As we discuss in section 12.4.4, there is no reason, in principle, why the data need to be randomly sampled at all. The null hypothesis says that the x's and the y's are indistinguishable from a random sample from the complete set of x's and y's together. We don't use the precise numbers — which depend upon the unknown distribution of the x's and y's — but only basic reasoning about the relative sizes of the numbers. Thus, if the x's and y's come from the same distribution, it is equally likely that  $x_i > y_i$  as that  $x_i < y_i$ .

### 12.4.1 Sign test

The idea of the sign test is quite straightforward. We wish to test the null hypothesis that paired data came from the same distribution. If that is the case, then which one of the two observations is the larger should be just like a coin flip. So we count up the number of times (out of  $n$  pairs) that the first observation in the pair is larger than the second, and compute the probability of getting that many heads in  $n$  coin flips. If that probability is below the chosen significance level  $\alpha$ , we reject the null hypothesis.

#### Schizophrenia study

Consider the schizophrenia study, discussed in section 11.5.2. We wish to test, at the 0.05 level, whether the schizophrenic twins have different brain measurements than the unaffected twins, against the null hypothesis that the measurements are really drawn from the same distribution. We focus now on the differences. Instead of looking at their values (which depends upon the underlying distribution) we look only at their signs, as indicated in Table 12.5. The idea is straightforward: Under the null hypothesis, the difference has an equal chance of being positive or negative, so the number of positive signs should be like the number of heads in fair coin flips. In this case, we have 14 + out of 15, which is obviously highly unlikely.

Table 12.5: Data from the Suddath [SCT<sup>+</sup>90] schizophrenia experiment. Hippocampus volumes in  $cm^3$ .

Unaffected	Schizophrenic	Difference	Sign
1.94	1.27	0.67	+
1.44	1.63	-0.19	-
1.56	1.47	0.09	+
1.58	1.39	0.19	+
2.06	1.93	0.13	+
1.66	1.26	0.40	+
1.75	1.71	0.04	+
1.77	1.67	0.10	+
1.78	1.28	0.50	+
1.92	1.85	0.07	+
1.25	1.02	0.23	+
1.93	1.34	0.59	+
2.04	2.02	0.02	+
1.62	1.59	0.03	+
2.08	1.97	0.11	+

Formally, we analyse this with a single-sample Z test, comparing the proportion of + to 0.5. We have  $n = 15$  trials, so the  $SE$  is  $\sqrt{0.5 \cdot 0.5 / 15} = 0.129$ , while the observed proportion of + is  $14/15 = 0.933$ . The Z statistic is

$$Z = \frac{0.933 - 0.5}{0.129} = 3.36,$$

which is far above the cutoff level 1.96 for rejecting the null hypothesis at the 0.05 level. In fact, the probability of getting such an extreme result (the so-called *p-value*) is less than 0.001.

What if we had ignored the pairing and applied the median test instead? We find that the median is 1.665. There are 9 schizophrenic and 6 unaffected twins with measures above the median, so  $p_u = 0.4$  and  $p_s = 0.6$ . The difference is small, as the standard error is  $SE = \sqrt{0.5 \cdot 0.5} \sqrt{1/15 + 1/15} = 0.18$ . We compute  $Z = 1.10$ , which does not allow us to reject the null hypothesis at any level. The rank-sum test also turns out to be too weak to reject the null hypothesis.

### 12.4.2 Breastfeeding study

We adopt an example from [vBFHL04], discussing a study by Brown and Hurlock on the effectiveness of three different methods of preparing breasts

for breastfeeding. Each mother treated one breast and left the other untreated, as a control. The two breasts were rated daily for level of discomfort, on a scale 1 to 4. Each method was used by 19 mothers, and the average difference between the treated and untreated breast for each of the 19 mothers who used the “toughening” treatment were:  $-0.525, 0.172, -0.577, 0.200, 0.040, -0.143, 0.043, 0.010, 0.000, -0.522, 0.007, -0.122, -0.040, 0.000, -0.100, 0.050, -0.575, 0.031, -0.060$ .

The original study performed a one-tailed t test at the 0.05 level of the null hypothesis that the true difference between treated and untreated breasts was 0: The cutoff is then  $-1.73$  (so we reject the null hypothesis on any value of  $T$  below  $-1.73$ ). We have  $\bar{x} = -0.11$ , and  $s_x = 0.25$ . We compute then  $T = (\bar{x} - 0)/(s_x/\sqrt{n}) = -1.95$ , leading us to reject the null. We should, however, be suspicious of this marginal result, which depends upon the choice of a one-tailed test: for a two-tailed test the cutoff would have been 2.10.

In addition, we note that the assumption of normality is drastically violated, as we see from the histogram of the observed values 12.4. To apply the sign test, we see that there are 8 positive and 9 negative values, which is as close to an average value as we could have, and so conclude that there is no evidence in the sign test of a difference between the treated and untreated breasts. (Formally, we could compute  $\hat{p} = 8/17 = 0.47$ , and  $Z = (0.47 - 0.50)/(0.5/\sqrt{17}) = -0.247$ , which is nowhere near the cutoff of 1.96 for the z test at the 0.05 level.)

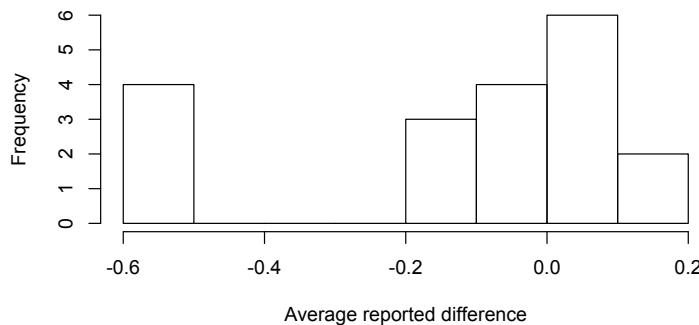


Figure 12.4: Histogram of average difference between treated and untreated breasts, for 19 subjects.

Historical note: This study formed part of the “rediscovery” of breastfeeding in the 1970s, after a generation of its being disparaged by the medical community. Their overall conclusion as that the traditional treatments were ineffective, the “toughening” marginally so. The emphasis on breastfeeding being fundamentally uncomfortable reflects the discomfort that the medical research community felt about nursing at the time.

### 12.4.3 Wilcoxon signed-rank test

As with the two-sample test in section 12.3.2, we can strengthen the paired-sample test by considering not just which number is bigger, but the relative ranks. The idea of the Wilcoxon (or *signed-rank*) test is that we might have about equal numbers of positive and negative values, but if the positive values are much bigger than the negative (or vice versa) that will still be evidence that the distributions are different. For instance, in the breastfeeding study, the t test produced a marginally significant result because several of the very large values are all negative.

The mechanics of the test are the same as for the two-sample rank-sum test, only the two samples are not the x’s and the y’s, but the positive and negative differences. In a first step, we rank the differences by their absolute values. Then, we carry out a rank-sum test on the positive and negative differences. To apply the Wilcoxon test, we first drop the two 0 values, and then rank the remaining 17 numbers by their absolute values:

Diff	0.007	0.010	0.031	0.040	-0.040	0.043	0.050	-0.060	-0.100
Rank	1	2	3	4.5	4.5	6	7	8	9
Diff	-0.122	-0.143	0.172	0.200	-0.522	-0.525	-0.575	-0.577	
Rank	10	11	12	13	14	15	16	17	

The ranks corresponding to positive values are 1, 2, 3 ,4.5, 6, 7, 12, 13, which sum to  $R_+ = 48.5$ , while the negative values have ranks 4.5,8,9,10,11,14,15,16,17, summing to  $R_- = 104.5$ . The Wilcoxon statistic is defined to be  $T = \min\{R_+, R_-\} = 48.5$ . We look on the appropriate table (given in Figure 12.5). We see that in order for the difference to be significant at the 0.05 level, we would need to have  $T \leq 34$ . Consequently, we still conclude that the effect of the treatment is not statistically significant.

n	P = 0.10	P = 0.05
5	2	-
6	2	0
7	3	2
8	5	3
9	8	5
10	10	8
11	14	10
12	17	13
13	21	17
14	26	21
15	30	25
16	36	29
17	41	34
18	47	40
19	53	46
20	60	52
21	67	58
22	75	65
23	83	73
24	91	81
25	100	89

Figure 12.5: Critical values for Wilcoxon test.

#### 12.4.4 The logic of non-parametric tests

One advantage of the non-parametric test is that it avoids the assumption that the means are sufficiently normal to fit the t distribution. In fact, though, we have presented reasonable evidence in section 11.5.3 that the sample means are quite close to normal. Is there any other reason to use a non-parametric test? Yes. The non-parametric test also avoids logically suspect assumptions which underly the parametric test.

As in section 11.6, there seems to be a logical hole in the application of the t test in the schizophrenia study of sections 11.5.2 and 11.5.3. Suppose we accept that the average of a random sample of 15 differences like the ones we've observed should have an approximately normal distribution, with mean equal to the population mean and variance equal to the population variance. Do we actually have such a random sample? The answer is, almost certainly, no. There is no population that these fifteen twin pairs were randomly sampled from. There is no register of all schizophrenia-discordant identical twin pairs from which these fifteen could have been randomly sampled.

We also cannot apply the logic of 11.6, which would say that everyone has

a “schizophrenic” and an “unaffected” measurement, and that we randomly decide whether we observe the one or the other. Is there some logic that we can apply? Yes. Imagine that some prankster had tampered with our data, randomly swapping the labels on the data, flipping a coin to decide which measurement would be called “schizophrenic” and which would be called “unaffected”. In other words, she randomly flips the signs on the differences between positive and negative. Our null hypothesis would say that we should not be able to recognise this tampering, because the two measurements are indistinguishable overall. What the sign test and the Wilcoxon test tell us is that our actual observations are not consistent with this hypothesis.

Note that the null hypothesis is now not tied to any probability model of the source of the actual data. Rather, it expresses our intuitive notion that “the difference is due to chance” in terms of a randomisation that could be performed. We have 15 cards, one for every twin pair in the study. One side has the measurement for the schizophrenic twin, the other has the measurement for the unaffected twin, but we have not indicated which is which. The null hypothesis says: The schizophrenic and the unaffected measurements have *the same distribution*, which means that we are just as likely to have either side of the card be the schizophrenic measurement. The differences we observe should be just like the differences we would observe if we labelled the sides schizophrenic and unaffected purely by flipping a coin. The signs test shows that this is not the case, forcing us to reject the null hypothesis. Statistical analysis depends on seeing the data we observe in the context of the entire array of equivalent observations that we could have made.

## Lecture 13

# Non-Parametric Tests Part II, Power of Tests

### 13.1 Kolmogorov-Smirnov Test

#### 13.1.1 Comparing a single sample to a distribution

In Figure 11.2 we compared some data — in Figure 11.2(c) these were 15 differences in brain measurements between schizophrenic and unaffected subjects, while in Figure 11.2(b) they were simulated data, from random resampling and averaging the original 15 — to a normal distribution. While these *Q-Q plots* seem to tell us something, it is hard to come to a definitive conclusion from them. Is the fit to the line close enough or not? It is this question that the Kolmogorov-Smirnov test is supposed to answer.

The basic setup of the Kolmogorov-Smirnov test is that we have some observations  $X_1, X_2, \dots, X_n$  which we think may have come from a given population (probability distribution)  $P$ . You can think of  $P$  as being a box with tickets in it, and the numbers representing the values that you might sample. We wish to test the null hypothesis

$$H_0 : \text{The samples came from } P$$

against the general alternative that the samples did not come from  $P$ . To do this, we need to create a test statistic whose distribution we know, and which will be big when the data are far away from a typical sample from the population  $P$ .

You already know one approach to this problem, using the  $\chi^2$  test. To do this, we split up the possible values into  $K$  ranges, and compare the

number of observations in each range with the number that would have been predicted. For instance, suppose we have 100 samples which we think should have come from a standard normal distribution. The data are given in Table 13.1. The first thing we might do is look at the mean and variance of the sample: In this case, the mean is  $-0.06$  and the sample variance  $1.06$ , which seems plausible. (A z test for the mean would not reject the null hypothesis of 0 mean, and the test for variance — which you have not learned — would be satisfied that the variance is 1.) We might notice that the largest value is  $3.08$ , and the minimum value is  $-3.68$ , which seem awfully large. We have to be careful, though, about scanning the data first, and then deciding what to test after the fact: This approach, sometimes called **data snooping**, can easily mislead, since every collection of data is likely to have *something* that seems wrong with it, purely by chance. (This is the problem of **multiple testing**, which we discuss further in section 14.3.)

-0.16	-0.68	-0.32	-0.85	0.89	-2.28	0.63	0.41	0.15	0.74
1.30	-0.13	0.80	-0.75	0.28	-1.00	0.14	-1.38	-0.04	-0.25
-0.17	1.29	0.47	-1.23	0.21	-0.04	0.07	-0.08	0.32	-0.17
0.13	-1.94	0.78	0.19	-0.12	-0.19	0.76	-1.48	-0.01	0.20
-1.97	-0.37	3.08	-0.40	0.80	0.01	1.32	-0.47	2.29	-0.26
-1.52	-0.06	-1.02	1.06	0.60	1.15	1.92	-0.06	-0.19	0.67
0.29	0.58	0.02	2.18	-0.04	-0.13	-0.79	-1.28	-1.41	-0.23
0.65	-0.26	-0.17	-1.53	-1.69	-1.60	0.09	-1.11	0.30	0.71
-0.88	-0.03	0.56	-3.68	2.40	0.62	0.52	-1.25	0.85	-0.09
-0.23	-1.16	0.22	-1.68	0.50	-0.35	-0.35	-0.33	-0.24	0.25

Table 13.1: Data possibly from standard normal distribution

The  $X^2$  statistic for this table is 7.90. This does not exceed the threshold of 9.49 for rejecting the null hypothesis at the 0.05 level.

There are two key problems with this approach:

- (1). We have thrown away some of the information that we had to begin with, by forcing the data into discrete categories. Thus, the power to reject the null hypothesis is less than it could have been. The bottom category, for instance, does not distinguish between  $-2$  and the actually observed extremely low observation  $-3.68$ .
- (2). We have to draw arbitrary boundaries between categories, and we may question whether the result of our significance test would have come

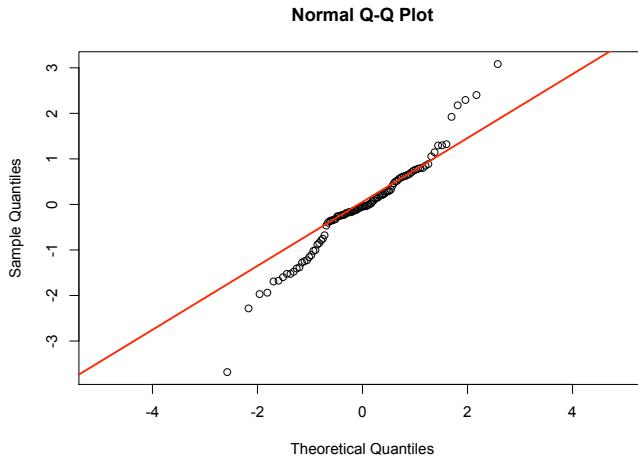
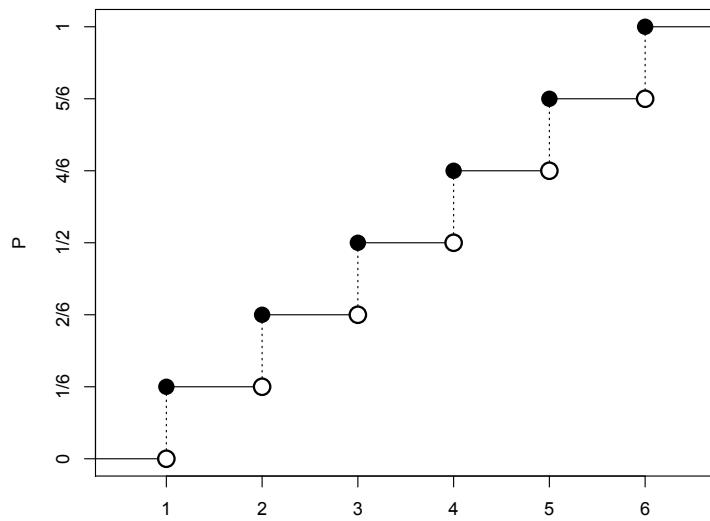


Figure 13.1: QQ plot of data from Table 13.1.

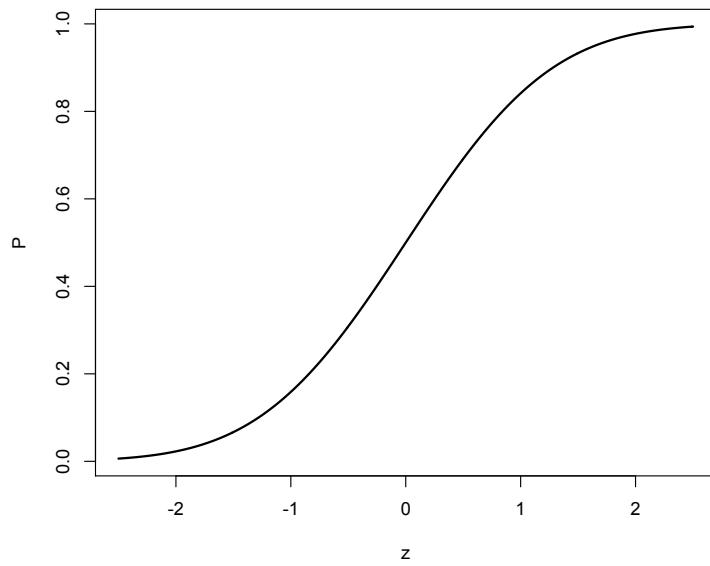
out differently if we had drawn the boundaries otherwise.

An alternative approach is to work directly from the intuition of figure 13.1: Our test statistic is effectively the maximum distance of the cumulative probability from the diagonal line. One important concept is the **distribution function** (or **cumulative distribution function**, or **cdf**), written  $F(x)$ . This is one way of describing a probability distribution. For every  $x$ , we define  $F(x)$  to be the probability of all values  $\leq x$ . Thus, if we are looking at the probability distribution of a single fair die roll, we get the cdf shown in Figure 13.2(a); the cdf of a normal distribution is in Figure 13.2(b). Note that the cdf always starts at 0, ends at 1, and has jumps when the distribution is discrete.

The **empirical distribution function**  $F_{obs}(x)$  of some data is simply the function that tells you what fraction of the data are below  $x$ . We show this for our normal (possibly) data in Figure 13.3(a). The “expected” distribution function  $F_{exp}(x)$  is the cdf predicted by the null hypothesis — in this case, the For each number  $x$ , we let  $F_{exp}(x)$  be the fraction of the numbers in our target population that are below  $x$ , and  $F_{obs}(x)$  the observed fraction below  $x$ . We then compute the Kolmogorov-Smirnov statistic, which is simply the maximum value of  $|F_{exp}(x) - F_{obs}(x)|$ . The practical approach is as follows: First, order the sample. We order our normal sample in Table 13.2(a) and give the corresponding normal probabilities (from the normal



(a) cdf for a fair die



(b) cdf for normal distribution

Figure 13.2: Examples of cumulative distribution functions

Category	Lower	Upper	Observed	Expected
1	$\infty$	-1.5	9	6.7
2	-1.5	-0.5	15	24.2
3	-0.5	0.5	49	38.3
4	0.5	1.5	22	24.2
5	1.5	$\infty$	5	6.7

Table 13.2:  $\chi^2$  table for data from Table 13.1, testing its fit to a standard normal distribution.

table) in Table 13.2(b). These probabilities need to be compared to the probabilities of the sample, which are just  $0.01, 0.02, \dots, 1.00$ . This procedure is represented graphically in Figure 13.3.

The Kolmogorov-Smirnov statistic is the maximum difference, shown in blue in Table 13.3. This is  $D_n = 0.092$ . For a test at the 0.05 significance level, we compare this to the critical value, which is

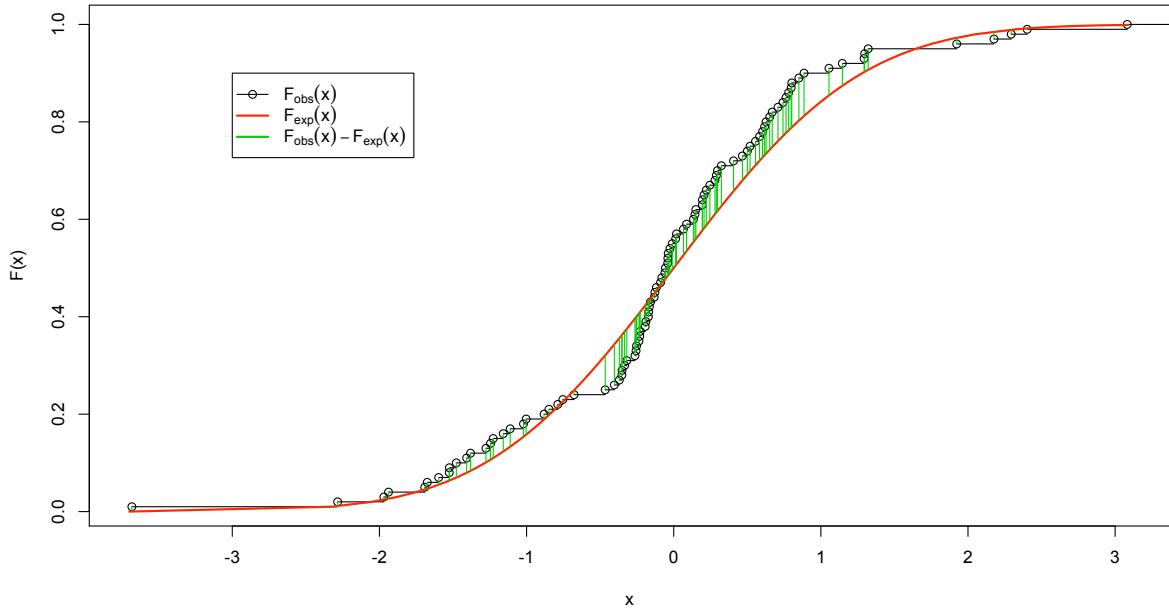
$$D_{crit} = \frac{1.36}{\sqrt{n}}$$

In this case, with  $n = 100$ , we get  $D_{crit} = 0.136$ . Since our observed  $D_n$  is smaller, we do not reject the null hypothesis.

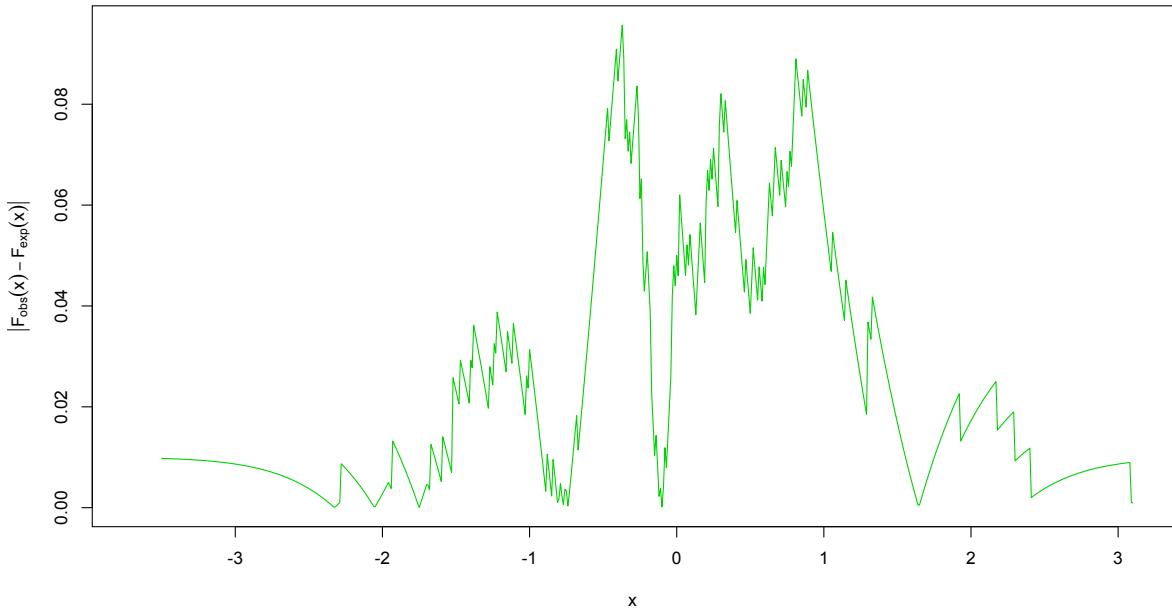
Of course, if you wish to compare the data to the normal distribution with mean  $\mu$  and variance  $\sigma^2$ , the easiest thing to do is to standardise: The hypothesis that  $(x_i)$  come from the  $N(\mu, \sigma^2)$  distribution is equivalent to saying that  $(x_i - \mu)/\sigma$  come from the standard  $N(0, 1)$  distribution. (Of course, if  $\mu$  and  $\sigma$  are estimated from the data, we get a Student t distribution in place of the standard normal.)

One final point: In point of fact, the data are unlikely to have come from a normal distribution. One way of seeing this is to look at the largest (negative) data point, which is  $-3.68$ . The probability of a sample from a standard normal distribution being at least this large is about 0.0002. In 100 observations, the probability of observing such a large value at least once is no more than 100 times as big, or 0.02. We could have a goodness-of-fit test based on the largest observation, which would reject the hypothesis that this sample came from a normal distribution. The Kolmogorov-Smirnov test, on the other hand, is indifferent to the size of the largest observation.

There are many different ways to test a complicated hypothesis, such as the equality of two distributions, because there are so many different ways



(a)  $F_{obs}$  shown in black circles, and  $F_{exp}$  (the normal distribution function) in red. The green segments show the difference between the two distribution functions.



(b) Plot of  $|F_{obs} - F_{exp}|$

Figure 13.3: Computing the Kolmogorov-Smirnov statistic for the data of Table 13.1.

(a) Ordered form of data from Table 13.1

-3.68	-2.28	-1.97	-1.94	-1.69	-1.68	-1.60	-1.53	-1.52	-1.48
-1.41	-1.38	-1.28	-1.25	-1.23	-1.16	-1.11	-1.02	-1.00	-0.88
-0.85	-0.79	-0.75	-0.68	-0.47	-0.40	-0.37	-0.35	-0.35	-0.33
-0.32	-0.26	-0.26	-0.25	-0.24	-0.23	-0.23	-0.19	-0.19	-0.17
-0.17	-0.17	-0.16	-0.13	-0.13	-0.12	-0.09	-0.08	-0.06	-0.06
-0.04	-0.04	-0.04	-0.03	-0.01	0.01	0.02	0.07	0.09	0.13
0.14	0.15	0.19	0.20	0.21	0.22	0.25	0.28	0.29	0.30
0.32	0.41	0.47	0.50	0.52	0.56	0.58	0.60	0.62	0.63
0.65	0.67	0.71	0.74	0.76	0.78	0.80	0.80	0.85	0.89
1.06	1.15	1.29	1.30	1.32	1.92	2.18	2.29	2.40	3.08

(b) Normal probabilities corresponding to Table 13.2(a)

0.000	0.011	0.024	0.026	0.045	0.047	0.055	0.064	0.064	0.070
0.080	0.084	0.101	0.107	0.110	0.123	0.133	0.154	0.158	0.189
0.198	0.215	0.226	0.249	0.321	0.343	0.356	0.362	0.363	0.369
0.375	0.396	0.399	0.400	0.407	0.409	0.410	0.423	0.425	0.432
0.432	0.434	0.437	0.447	0.449	0.453	0.464	0.468	0.476	0.477
0.484	0.484	0.485	0.490	0.496	0.505	0.508	0.526	0.535	0.553
0.557	0.560	0.577	0.577	0.582	0.588	0.597	0.610	0.614	0.617
0.627	0.658	0.680	0.692	0.698	0.711	0.720	0.727	0.732	0.735
0.743	0.748	0.761	0.771	0.777	0.783	0.788	0.789	0.803	0.812
0.854	0.874	0.902	0.903	0.907	0.973	0.985	0.989	0.992	0.999

(c) Difference between entry # $i$  in Table 13.2(b) and  $i/100$ . Largest value shown blue.

0.010	0.009	0.006	0.014	0.004	0.014	0.015	0.017	0.026	0.031
0.031	0.036	0.030	0.034	0.041	0.037	0.037	0.026	0.031	0.011
0.012	0.005	0.003	0.008	0.069	0.085	0.086	0.083	0.073	0.071
0.064	0.077	0.067	0.061	0.055	0.049	0.039	0.045	0.035	0.033
0.023	0.013	0.006	0.008	0.002	0.008	0.006	0.012	0.014	0.024
0.026	0.036	0.046	0.052	0.054	0.056	0.062	0.052	0.054	0.048
0.054	0.060	0.055	0.061	0.067	0.073	0.071	0.070	0.076	0.082
0.084	0.061	0.049	0.049	0.052	0.048	0.051	0.054	0.058	0.064
0.068	0.071	0.069	0.070	0.074	0.078	0.082	0.092	0.088	0.087
0.055	0.045	0.029	0.037	0.043	0.013	0.015	0.009	0.002	0.001

Table 13.3: Computing the Kolmogorov-Smirnov statistic for testing the fit of data to the standard normal distribution.

for distributions to differ. We need to choose a test that is sensitive to the differences that we expect (or fear) to find between reality and our null hypothesis. We already discussed this in the context of one- and two-tailed

t and Z tests. If the null hypothesis says  $\mu = \mu_0$ , and the alternative is that  $\mu > \mu_0$ , then we can increase the power of our test against this alternative by taking a one-sided alternative. On the other hand, if reality is that  $\mu < \mu_0$  then the test will have essentially no power at all. Similarly, if we think the reality is that the distributions of  $X$  and  $Y$  differ on some scattered intervals, we might opt for a  $\chi^2$  test.

### 13.1.2 Comparing two samples: Continuous distributions

Suppose an archaeologist has measured the distances of various settlements from an important cult site at two different periods — let us say,  $X_1, \dots, X_{10}$  are the distances (in km) for the early settlements, and  $Y_1, \dots, Y_{12}$  are the distances for the late settlements. She is interested to know whether there has been a change in the settlement pattern, and believes that these settlements represent a random sample from a large number of such settlements in these two periods. The measurements are:

$$\textcolor{red}{X} : 5.4, 19.3, 15.8, 4.9, 0.7, 4.9, 8.5, 23.1, 16.7, 2.0$$

$$\textcolor{green}{Y} : 27.1, 28.4, 5.6, 29.0, 29.8, 26.1, 14.4, 14.9, 29.3, 18.0, 0.4, 23.3.$$

We put these values in order to compute the cdf (which we also plot in Figure 13.4).

	1.2	1.4	1.9	3.7	4.4	4.8	5.6	6.5	6.6	6.9	9.2	9.7	10.4	10.6	17.3	19.3	21.1	28.4
$F_x$	0.1	0.2	0.3	0.4	0.5	0.6	0.6	0.6	0.6	0.6	0.7	0.7	0.7	0.7	0.8	0.8	0.9	1.0
$F_y$	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.4	0.5	0.6	0.6	0.8	0.9	0.9	1.0	1.0	1.0

Table 13.4: Archaeology (imaginary) data, tabulation of cdfs.

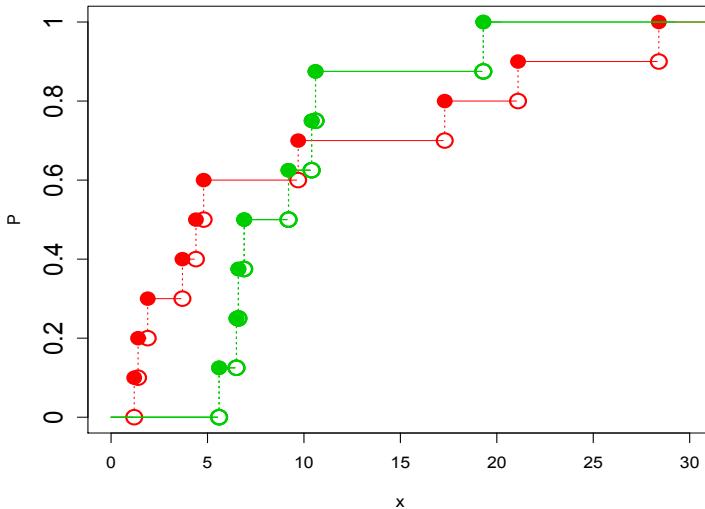


Figure 13.4: Cumulative distribution functions computed from “archaeology” data, as tabulated in Table 13.4.

The Kolmogorov-Smirnov statistic is  $D = 0.6$ . We estimate the critical value from this formula:

Kolmogorov-Smirnov Critical value:  

$$D_{\text{crit},0.05} = 1.36 \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 1.36 \sqrt{\frac{n_x+n_y}{n_x n_y}}.$$

With  $n_x = 10$  and  $n_y = 8$ , this yields  $D_{\text{crit},0.05} = 0.645$ , so that we just miss rejecting the null hypothesis. (As it happens, the approximate critical value is too high here. The true critical value is exactly 0.6. A table of exact critical values is available at [http://www.soest.hawaii.edu/wessel/courses/gg313/Critical\\_KS.pdf](http://www.soest.hawaii.edu/wessel/courses/gg313/Critical_KS.pdf).)

### 13.1.3 Comparing two samples: Discrete samples

The Kolmogorov-Smirnov test is designed to deal with samples from a continuous distribution without having to make arbitrary partitions. It is not appropriate for samples from discrete distributions, or when the data are

presented in discrete categories. Nonetheless, it is often used for such applications. We present here one example.

One method that anthropologists use to study the health and lives of ancient populations, is to estimate the age at death from skeletons found in gravesites, and compare the distribution of ages. The paper [Lov71] compares the age distribution of remains found at two different sites in Virginia, called Clarksville and Tollifero. The data are tabulated in Table 13.5.

Table 13.5: Ages at death for skeletons found at two Virginia sites, as described in [Lov71].

Age range	Clarksville	Tollifero
0–3	6	13
4–6	0	6
7–12	2	4
13–17	1	9
18–20	0	2
21–35	12	29
35–45	15	8
45–55	1	8
55+	0	0
n	37	79

The authors used the Kolmogorov-Smirnov test to compare two different empirical distributions, in exactly the same way as to compare an empirical distribution with a theoretical distribution. We compute the cumulative distributions of the two distributions, just as before — sometimes these are denoted  $F_1$  and  $F_2$  — and we look for the maximum difference between them. The calculations are shown in Table 13.6. We see that the maximum difference is at age class 21–35, with  $D = 0.23$ .

It remains to interpret the result. We are testing the null hypothesis that the Clarksville and Tollifero ages were sampled from the same age distribution, against the alternative that they were sampled from different distributions. With  $n_x = 37$  and  $n_y = 79$ , we get  $D_{crit,0.05} = 0.27$ . Since the observed  $D$  is smaller than this, we do not reject the null hypothesis.

This is a common application, but it has to be said that this doesn't

Table 13.6: Calculations for Kolmogorov-Smirnov test on data from Table 13.5.

Age range	Cumulative Distrib.		
	Clarksville	Tollifero	Difference
3	0.16	0.16	0
6	0.16	0.24	0.08
12	0.22	0.29	0.07
17	0.24	0.41	0.17
20	0.24	0.43	0.19
35	0.57	0.8	0.23
45	0.97	0.9	0.07
55	1	1	0

entirely make sense. What we have done is effectively to take the maximum difference not over all possible points in the distribution, but only at eight specially chosen points. This inevitably makes the maximum smaller. The result is to make it harder to reject the null hypothesis, so our significance level is too high. We should compensate by lowering the critical value.

### 13.1.4 Comparing tests to compare distributions

Note that the Kolmogorov-Smirnov test differs from  $\chi^2$  in another important way: The chi-squared statistic doesn't care what order the categories come in, while order is crucial to the Kolmogorov-Smirnov statistic. This may be good or bad, depending on the kinds of deviation from the null hypothesis you think are important. (Thus, in our example here, Kolmogorov-Smirnov would record a larger deviation from the null hypothesis if the Clarksville site had lower mortality at all juvenile age classes, than if it had lower mortality in age classes 0–3 and 35–55, and higher mortality elsewhere. There are alternative tests

## 13.2 Power of a test

What is a hypothesis test? We start with a null hypothesis  $H_0$ , which for our purposes is a distribution that the observed data may have come from, and a level  $\alpha$ , and we wish to determine whether the data could have had

at least a chance  $\alpha$  of being observed if  $H_0$  were true. A test of  $H_0$  at level  $\alpha$  is a *critical region*  $\mathcal{R}$ , a set of possible observations for which you would reject  $H_0$ , which must satisfy  $P_0\{\text{data in } \mathcal{R}\} = \alpha$ . That is, if  $H_0$  is true, the probability is just  $\alpha$  that we will reject the null hypothesis. We reject  $H_0$  if the observations are such as would be unlikely if  $H_0$  were true.

How about this for a simple hypothesis test: Have your computer random number generator pick a random number  $Y$  uniformly between 0 and 1. No matter what the data are, you reject the null hypothesis if  $Y < \alpha$ . This satisfies the rule: The probability of rejecting the null hypothesis is always  $\alpha$ . Unfortunately, the probability of rejecting the null hypothesis is also only  $\alpha$  if the alternative (any alternative) is true.

Of course, a good test isn't just any procedure that satisfies the probability rule. We also want it to have a higher probability of rejecting the null hypothesis when the null is false. This probability is called the **power** of the test, and is customarily denoted  $1 - \beta$ . Here  $\beta$  is the probability of making a Type II Error: not rejecting the null hypothesis, even though it is false. All things being equal, a test with higher power is preferred, and if the power is too low the experiment is not worth doing at all. This is why power computations ought to take place at a very early stage of planning an experiment.

Decision \ Truth	$H_0$ True	$H_0$ False
Don't Reject $H_0$	Correct (Prob. $1 - \alpha$ )	Type II Error (Prob. $= \beta$ )
Reject $H_0$	Type I Error (Prob. = level $= \alpha$ )	Correct (Prob. = Power $= 1 - \beta$ )

### 13.2.1 Computing power

Power depends on the alternative: It is always the power to reject the null, given that a specific alternative is actually the case. Consider, for instance, the following idealised experiment: We make measurements  $X_1, \dots, X_{100}$  of a quantity that is normally distributed with unknown mean  $\mu$  and known variance 1. The null hypothesis is that  $\mu = \mu_0 = 0$ , and we will test it with a two-sided Z test at the 0.05 level, against a simple alternative  $\mu = \mu_{alt}$ . That is, we assume

What is the power? Once we have the data, we compute  $\bar{x}$ , and then  $z = \bar{x}/0.1$ . We reject the null hypothesis if  $z > 1.96$  or  $z < -1.96$ . The power is the probability that this happens. What is this probability? It is the same

as the probability that  $\bar{X} > 0.196$  or  $\bar{X} < -0.196$ . We know that  $\bar{X}$  has  $N(\mu, 0.1)$  distribution. If we standardise this, we see that  $Z' := 10(\bar{X} - \mu)$  is standard normal. A very elementary probability computation shows us that

$$\begin{aligned} P\{\bar{X} < -0.196\} &= P\{\bar{X} - \mu < -0.196 - \mu\} = P\{Z' < -1.96 - 10\mu\} \\ P\{\bar{X} > 0.196\} &= P\{\bar{X} - \mu > 0.196 - \mu\} = P\{Z' > 1.96 - 10\mu\} \end{aligned}$$

Thus

$$\text{Power} = \Phi(-1.96 - 10\mu_{alt}) + 1 - \Phi(1.96 - 10\mu_{alt}), \quad (13.1)$$

where  $\Phi$  is the number we look up on the normal table.

More generally, suppose we have  $n$  measures of a quantity with unknown distribution, unknown mean  $\mu$ , but known variance  $\sigma^2$ , and we wish to use a two-tailed Z test for the null hypothesis  $\mu = \mu_0$  against the alternative  $\mu = \mu_{alt}$ . The same computation as above shows that

$$\begin{aligned} \text{Power} &= \Phi\left(-z - \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu_{alt})\right) \\ &\quad + 1 - \Phi\left(z - \frac{\sqrt{n}}{\sigma}(\mu_0 - \mu_{alt})\right). \end{aligned} \quad (13.2)$$

To put this differently, the power is the probability of getting a decisive result (rejecting the null hypothesis) if the hidden real mean is  $\mu_{alt}$ .

You don't need to memorise this formula, but it is worth paying attention to the qualitative behaviour of the power. Figure 13.5 shows the power for a range of alternative numbers of samples, and levels of test. Note that increasing the number of samples increases the power, while lowering the level of the test decreases the power. Note, too, that the power approaches the level of the test, as the alternative  $\mu_{alt}$  approaches the null  $\mu_0$ .

### 13.2.2 Computing trial sizes

Suppose now that you are planning an experiment to test a new blood-pressure medication. Medication A has been found to lower blood pressure by 10 mmHg, with an SD of 8 mmHg, and we want to test it against the new medication B. We will recruit some number of subjects, randomly divide them into control and experiment groups; the control group receives A, the experiment group receives B. At the end we will perform a Z test at the 0.05 level to decide whether B really lowered the subjects' blood pressure more than A; that is, we will let  $\mu_A$  and  $\mu_B$  be the true mean effect of drugs A

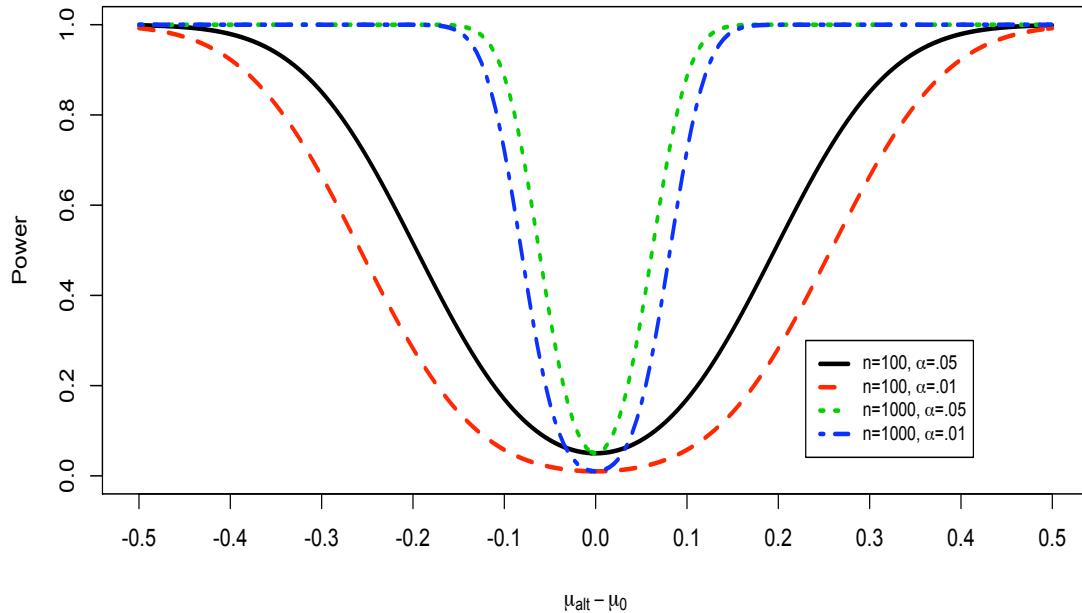


Figure 13.5: Power for Z test with different gaps between the null and alternative hypotheses for the mean, for given sizes of study ( $n$ ) and significance levels  $\alpha$ .

and B respectively on blood pressure, and test the null hypothesis  $\mu_A = \mu_B$  against the two-tailed alternative  $\mu_A \neq \mu_B$ , or the one-tailed alternative  $\mu_A < \mu_B$ . Following the discussion of section 11.6, we can act as though these were simple random samples of size  $n/2$  from one population receiving A and another receiving B, with the test statistic

$$Z = \frac{\bar{b} - \bar{a}}{\sigma \sqrt{\frac{1}{n/2} + \frac{1}{n/2}}}.$$

Is it worth doing this experiment? Given our resources, there is a limited number of subjects we can afford to recruit, and that determines the power of the test we do at the end. Our estimate of the power also depends, of course, on what we think the difference between the two population means

is. We see from the formula (13.2) that if the gap between  $\mu_A$  and  $\mu_B$  becomes half as big, we need 4 times as many subjects to keep the same power. If the power is only 0.2, say, then it is hardly worth starting in on the experiment, since the result we get is unlikely to be conclusive.

Figure 13.6 shows the power for experiments where the true experimental effect (the difference between  $\mu_A$  and  $\mu_B$ ) is 10 mmHg, 5 mmHg, and 1 mmHg), performing one-tailed and two-tailed significance tests at the 0.05 level.

Notice that the one-tailed test is always more powerful when  $\mu_B - \mu_A$  is on the right side ( $\mu_B$  bigger than  $\mu_A$ , so B is superior), but essentially 0 when B is inferior; the power of the two-tailed test is symmetric. If we are interested to discover only evidence that B is superior, then the one-tailed test obviously makes more sense.

Suppose now that we have sufficient funding to enroll 50 subjects in our study, and we think the study would be worth doing only if we have at least an 80% chance of finding a significant positive result. In that case, we see from Figure 13.6(b) that we should drop the project unless we expect the difference in average effects to be at least 5 mmHg. On the other hand, if we can afford 200 subjects, we can justify hunting for an effect only half as big, namely 2.5 mmHg. With 1000 subjects we have a good chance of detecting a difference between the drugs as small as 1 mmHg. On the other hand, with only 10 subjects we would be unlikely to find the difference to be statistically significant, even if the true difference is quite large.

**Important lesson:** The difference between a statistically significant result and a non-significant result may be just the size of the sample. Even an insignificant difference in the usual sense (i.e., tiny) has a high probability (power) of producing a statistically significant result if the sample size is large enough.

### 13.2.3 Power and non-parametric tests

The cost of dispensing with questionable assumptions is to reduce the power to reject a false null hypothesis in cases where the assumptions do hold. To see this, we consider one very basic example: We observe 10 samples from a normal distribution with mean  $\mu$  and variance 1, which is unknown. We do not know  $\mu$  either, so we wish to test the null hypothesis  $H_0 : \mu = 0$  against the alternative hypothesis that  $\mu \neq 0$ . We perform our test at the 0.05 level.

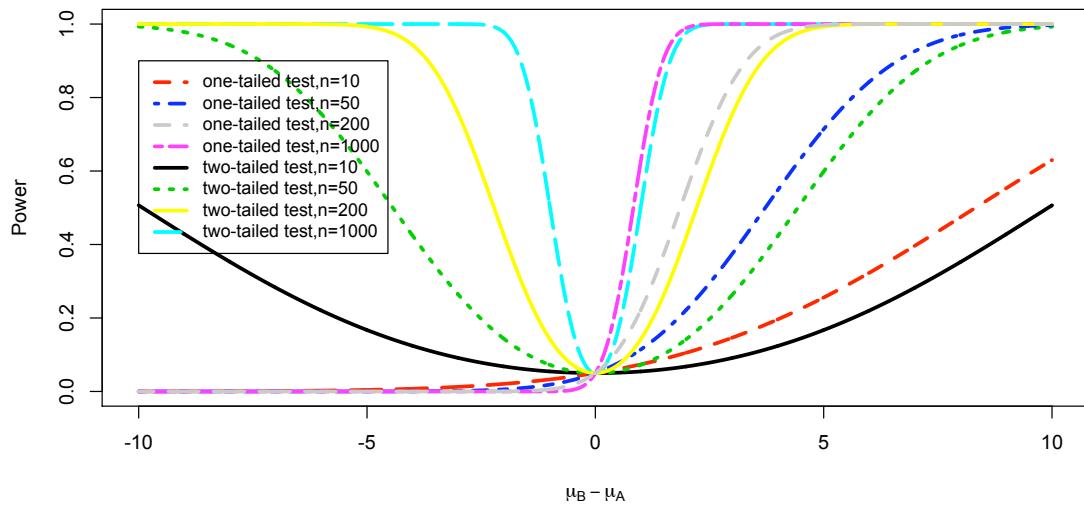
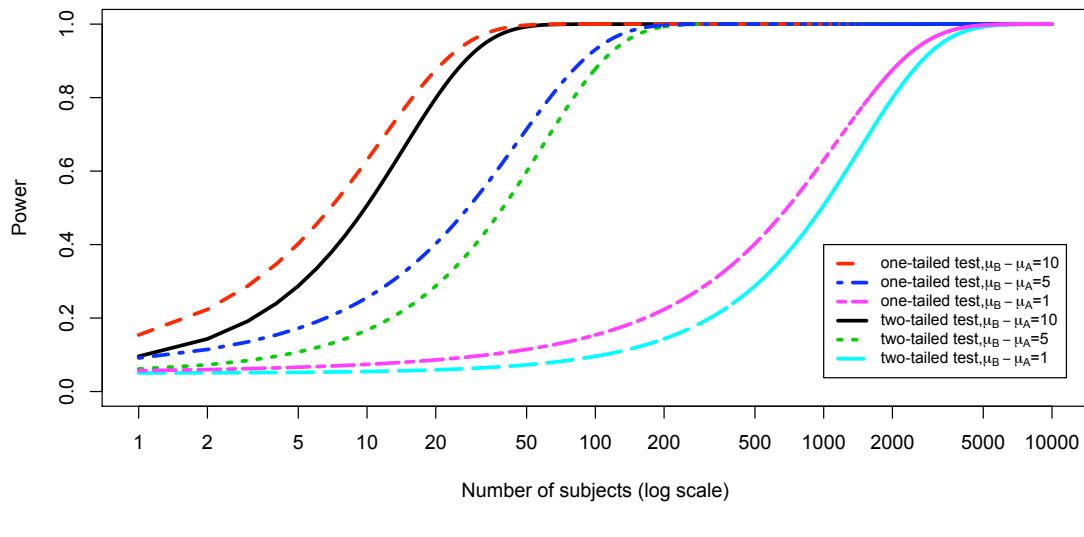


Figure 13.6: Power of the BP experiment, depending on number of subjects and true difference between the two means.

In figure 13.7 we show a plot of the probability (estimated from simulations) of rejecting the null hypothesis, as a function of the true (unobserved) mean. We compare the two-tailed t test with the median test and the rank-sum test. Notice that the median test performs far worse than the others, but that the Mann-Whitney test is only slightly less powerful than the t test, despite being far more general in its assumptions.

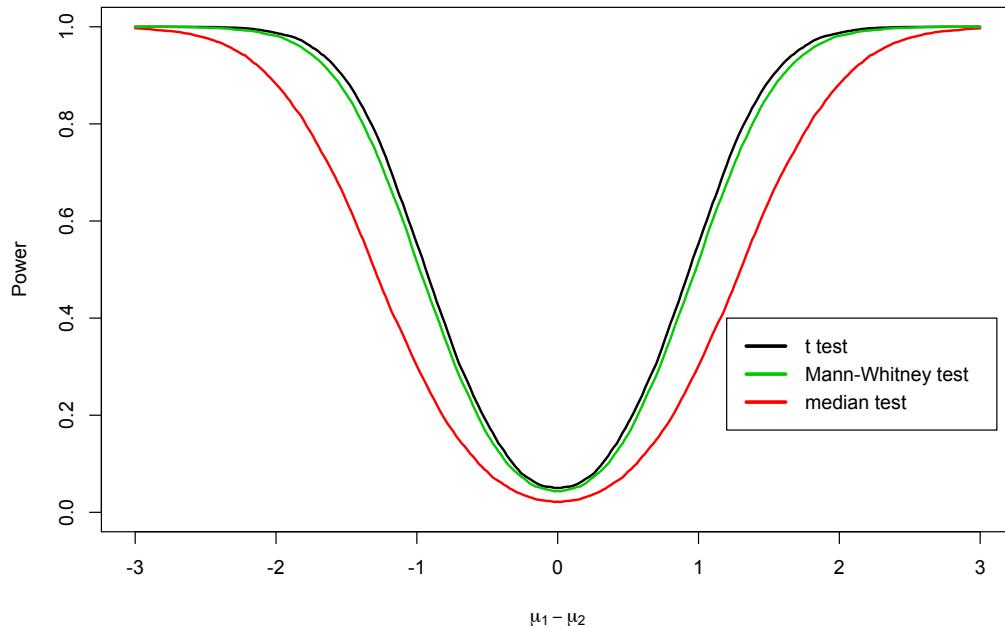


Figure 13.7: Estimated power for three different tests, where the underlying distributions are normal with variance 1, as a function of the true difference in means. The test is based on ten samples from each distribution.



# Lecture 14

## ANOVA and the F test

### 14.1 Example: Breastfeeding and intelligence

A study was carried out of the relationship between duration of breastfeeding and adult intelligence. The subjects were part of the Copenhagen Perinatal Cohort, 9125 individuals born at the Copenhagen University Hospital between October 1959 and December 1961. As reported in [MMSR02], a subset of the cohort was contacted for follow-up as adults (between ages 20 and 34). 983 subjects completed the Danish version of the Wechsler Adult Intelligence Scale (WAIS).

Table 14.1 shows the average scores for 3 tests, for 5 classes of breastfeeding duration. Look first at the rows marked “Unadjusted mean”. We notice immediately that 1) Longer breastfeeding was associated with higher mean intelligence scores, but 2) The longest breastfeeding (more than 9 months) was associated with lower mean scores. We ask whether either or both of these associations is reliably linked to the duration of breastfeeding, or whether they could be due to chance.

### 14.2 Digression: Confounding and the “adjusted means”

Before we can distinguish between these two possibilities — causation or chance association — we need to address another possibility: **confounding**. Suppose, for instance, that mothers who smoke are less likely to breastfeed. Since mother’s smoking is known to reduce the child’s IQ scores, this would produce higher IQ scores for the breastfed babies, irrespective of any causal influence of the milk. The gold standard for eliminating confounding is

Table 14.1: Intelligence scores (WAIS) by duration of breastfeeding.

Test	N	Duration of Breastfeeding (months)				
		≤ 1	2-3	4-6	7-9	> 9
Verbal IQ	Unadjusted mean	98.2	101.7	104.0	108.2	102.3
	SD	16.0	14.9	15.7	13.3	15.2
	Adjusted Mean	99.7	102.3	102.7	105.7	103.0
Performance IQ	Unadjusted mean	98.5	100.5	101.8	106.3	102.6
	SD	15.8	15.2	15.6	13.9	14.9
	Adjusted Mean	99.1	100.6	101.3	105.1	104.4
Full Scale IQ	Unadjusted mean	98.1	101.3	103.3	108.2	102.8
	SD	15.9	15.2	15.7	13.1	14.4
	Adjusted Mean	99.4	101.7	102.3	106.0	104.0

the double-blind random controlled experiment. Subjects are assigned at random to receive the treatment or not, so that the only difference between the two groups is whether they received the treatment. (“Double blind” refers to the use of protocols that keep the subjects and the experimenters from knowing who has received the treatment and who is a control. Without blinding, the two groups would differ in their knowledge of having received the treatment or not. We then might be unable to distinguish between effects which are actually due to the treatment, and those that come from *believing* you have received the treatment — in particular, the so-called **placebo effects**.)

Of course, it is usually neither possible nor ethical to randomly assign babies to different feeding regimens. What we have here is an observational study. The next best solution is then to try to remove the confounding. In this case, the researchers looked at all the factors that they might expect to have an effect on adult intelligence — maternal smoking, maternal height, parents’ income, infant’s birthweight, and so on — and adjusted the scores for each category to compensate for a preponderance of characteristics that might be expected to raise or lower IQ in that category, regardless of infant nutrition. Thus, we see that the first and last categories both had their means adjusted substantially upward, which must mean that the infants

who were nursed more than 9 months and those nursed less than 1 month both had, on average, characteristics (whether their own or their mothers') that would seem to predispose them to lower IQ. For the rest of this chapter we will work with the adjusted means.

The statistical technique for doing this, called *multiple regression*, is outside the scope of this course, but it is fairly straightforward, and most textbooks on statistical methods that go beyond the most basic techniques will describe it. Modern statistical software makes it particularly easy to adjust data with multiple regression.

## 14.3 Multiple comparisons

Let us consider the adjusted Full Scale IQ scores. We wish to determine whether the scores of individuals with the same breastfeeding class might have come from the same distribution, with the differences being solely due to random variation.

### 14.3.1 Discretisation and the $\chi^2$ test

One approach would be to group the IQ scores into groups — low, medium, and high, say. We would then have an incidence table. If these were categorical data — proportions of subjects in each breastfeeding class who scored “high” and “low”, for instance — we could produce an incidence table such as that in Table 14.2. (The data shown here are purely invented, for illustrative purposes.) You have learned how to analyse such a table to determine whether the vertical categories (IQ score) are independent of the horizontal categories (duration of breastfeeding), using the  $\chi^2$  test.

The problem with this approach is self-evident: We have thrown away some of the information that we had to begin with, by forcing the data into discrete categories. Thus, the power to reject the null hypothesis is less than it could have been. Furthermore, we have to draw arbitrary boundaries between categories, and we may question whether the result of our significance test would have come out differently if we had drawn the boundaries otherwise. (These are the same problems, you may recall, that led us to prefer the Kolmogorov-Smirnov test over  $\chi^2$ . The  $\chi^2$  test has the virtue of being wonderfully general, but it is often not quite the best choice.)

Table 14.2: Hypothetical incidence table, if IQ data were categorised into low, medium, and high

Full IQ score	Breastfeeding months				
	≤ 1	2-3	4-6	7-9	> 9
high	100	115	120	40	9
medium	72	85	69	35	9
low	100	115	80	29	5

### 14.3.2 Multiple t tests

Alternatively, we can compare the mean IQ scores between two different breastfeeding categories, using the t test — effectively, this is the z test, since the number of degrees of freedom is so large, but we still need to pool the variance, because one of the categories has a fairly small number of samples. (The large number of samples also allows us to be reasonably confident in treating the mean as normally distributed, as discussed in section 7.4.) For instance, suppose we wish to compare the children breastfed less than 1 month with those breastfed more than 9 months. We want to test for equality of means, at the 0.05 level.

We compute the pooled standard deviation by

$$s_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} = \sqrt{\frac{271 \cdot 15.9^2 + 22 \cdot 14.4^2}{293}} = 15.8,$$

and the standard error by

$$SE_{\text{diff}} = s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = 3.43.$$

This yields a t statistic of

$$t = \frac{\bar{x} - \bar{y}}{SE_{\text{diff}}} = \frac{-4.6}{3.42} = -1.34.$$

Since the cutoff is 1.96, we do not reject the null hypothesis.

If we repeat this test for all 10 pairs of categories, we get the results shown in Table 14.3. We see that 4 out of the 10 pairwise comparisons show statistically significant differences. But what story are these telling together? Remember that if the null hypothesis were true — if the population

means were in fact all the same — 1 out of 20 comparisons should yield a statistically significant difference at the 0.05 level. How many statistically significant differences do we need before we can reject the overall null hypothesis of identical population means? And what if none of the differences were individually significant, but they all pointed in the same direction?

Table 14.3: Pairwise t statistics for comparing all 10 pairs of categories. Those that exceed the significance threshold for the 0.05 level are shown in red.

	2-3	4-6	7-9	> 9
≤ 1	-1.78	<b>-2.14</b>	<b>-3.78</b>	-1.34
2-3		-0.47	<b>-2.58</b>	-0.70
4-6			<b>-2.14</b>	-0.50
7-9				0.65

## 14.4 The F test

We will see in lecture 15 how to treat the covariate — the duration of breastfeeding — as a *quantitative* rather than *categorical* variable. That is, how to measure the effect (if any) per unit time breastfeeding. Here we concern ourselves only with the question: Is there a nonrandom difference between the mean intelligence in the different categories? As we discussed in section 14.3, we want to reduce the question to a single test.

What should the test statistic look like? Fundamentally, a test statistic should have two properties: 1) It measures significant deviation from the null hypothesis; that is, one can recognise from the test statistic whether the null hypothesis has been violated substantially. 2) We can compute the distribution of the statistic under the null hypothesis.

### 14.4.1 General approach

Suppose we have independent samples from  $K$  different normal distributions, with means  $\mu_1, \dots, \mu_K$  and variance  $\sigma^2$  (so the variances are all the same). We call these  $K$  groups **levels** (or sometimes **treatments**). We have  $n_i$  samples from distribution  $i$ , which we denote  $X_{k1}, X_{k2}, \dots, X_{kn_k}$ . The goal

is to determine from these samples whether the  $K$  **treatment effects**  $\mu_k$  could be all equal.

We let  $N = \sum_{k=1}^K n_k$  be the total number of observations. The average of all the observations is  $\bar{X}$ , while the average within level  $i$  is  $\bar{X}_i$ :

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}.$$

The idea of analysis of variance (ANOVA) is that under the null hypothesis, which says that the observations from different levels really all are coming from the same distribution, the observations should be about as far (on average) from their own level mean as they are from the overall mean of the whole sample; but if the means are different, observations should be closer to their level mean than they are to the overall mean.

We define the **Between Groups Sum of Squares**, or **BSS**, to be the total square difference of the group means from the overall mean; and the **Error Sum of Squares**, or **ESS**, to be the total squared difference of the samples from the means of their own groups. (The term “error” refers to a context in which the samples can all be thought of as measures of the same quantity, and the variation among the measurements represents random error; this piece is also called the Within-Group Sum of Squares.) And then there is the **Total Sum of Squares**, or **TSS**, which is simply the total square difference of the samples from the overall mean, if we treat them as one sample.

$BSS = \sum_{i=1}^K n_i (\bar{X}_i - \bar{X})^2;$
$ESS = \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$
$= \sum_{i=1}^K (n_i - 1)s_i^2$ where $s_i$ is the SD of observations in level $i$ .
$TSS = \sum_{i,j} (X_{ij} - \bar{X})^2$
$= (N - 1)s^2$ , where $s$ is the sample SD of all observations together.

The initials **BMS** and **EMS** stand for **Between Groups Mean Squares** and **Error Mean Squares** respectively.

The **analysis of variance (ANOVA)** is based on two mathematical facts. The first is the identity  $TSS = ESS + BSS$ . In other words, all the variability among the data can be divided into two pieces: The variability within groups, and the variability among the means of different groups. Our goal is to evaluate the apportionment, to decide if there is “too much” between group variability to be purely due to chance.

Of course,  $BSS$  and  $ESS$  involve different numbers of observations in their sums, so we need to normalise them. We define

$$BMS = \frac{BSS}{K - 1} \quad EMS = \frac{ESS}{N - K}.$$

This brings us to the second mathematical fact: if the null hypothesis is true, then  $EMS$  and  $BMS$  are both estimates for  $\sigma^2$ . On the other hand, interesting deviations from the null hypothesis — in particular, where the populations have different means — would be expected to increase  $BMS$  relative to  $EMS$ . This leads us to define the deviation from the null hypothesis as the ratio of these two quantities:

$$F = \frac{BMS}{EMS} = \frac{N - K}{K - 1} \cdot \frac{BSS}{ESS}.$$

We reject the null hypothesis when  $F$  is too large: That is, if we obtain a value  $f$  such that  $P\{F \geq f\}$  is below the significance level of the test.

Table 14.4: Tabular representation of the computation of the F statistic.

	SS	d.f.	MS	F
Between Treatments	BSS (A)	$K - 1$ (B)	BMS ( $X = A/B$ )	$X/Y$
Errors (Within Treatments)	ESS (C)	$N - K$ (D)	EMS ( $Y = C/D$ )	
Total	TSS	$N - 1$		

Under the null hypothesis, the F statistic computed in this way has a known distribution, called the F distribution with  $(K - 1, N - K)$  degrees of freedom. We show the density of  $F$  for  $K = 5$  different treatments and different values of  $N$  in Figure 14.1.

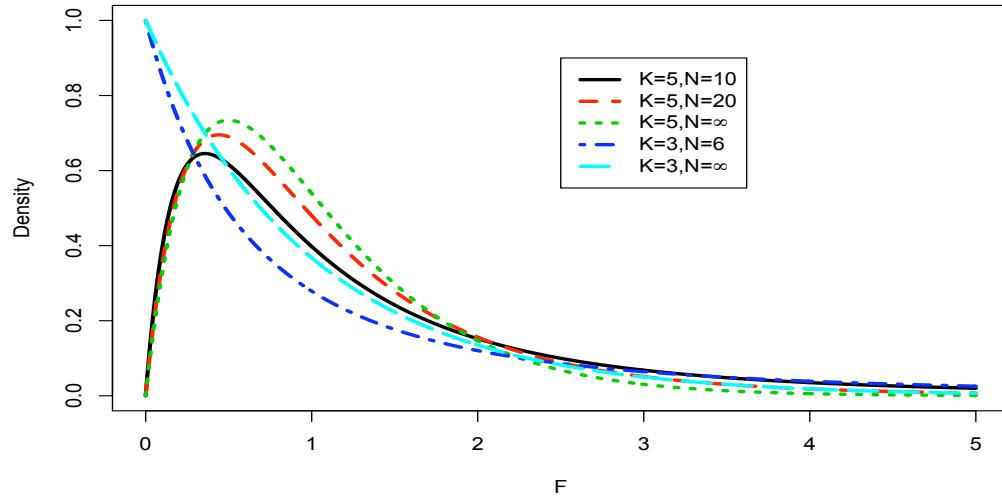


Figure 14.1: Density of  $F$  distribution for different values of  $K$  and  $N$ .

#### 14.4.2 The breastfeeding study: ANOVA analysis

In this case, since we do not know the individual observations, we cannot compute TSS directly. We compute

$$\begin{aligned}
 ESS &= \sum_{k=1}^5 (n_k - 1)s_k^2 \\
 &= 271 \cdot 15.9^2 + 304 \cdot 15.2^2 + 268 \cdot 15.7^2 + 104 \cdot 13.1^2 + 23 \cdot 14.4^2 \\
 &= 227000; \\
 BSS &= \sum_{k=1}^5 n_k(x_{k\cdot} - \bar{x})^2 \\
 &= 272 \cdot (99.4 - 101.7)^2 + 305 \cdot (101.7 - 101.7)^2 + 269 \cdot (102.3 - 101.7)^2 \\
 &\quad + 104 \cdot (106.0 - 101.7)^2 + 23 \cdot (104.0 - 101.7)^2 \\
 &= 3597.
 \end{aligned}$$

We complete the computation in Table 14.5, obtaining  $F = 3.81$ . The numbers of degrees of freedom are  $(4, 968)$ . The table in the official booklet is

quite small — after all, there is one distribution for each pair of integers. The table gives only the cutoff only for select values of  $(d_1, d_2)$  at the 0.05 level. For parameters in between one needs to interpolate, and for parameters above the maximum we go to the row or column marked  $\infty$ . Looking on the table in Figure 14.2, we see that the cutoff for  $F(4, \infty)$  is 2.37. Using a computer, we can compute that the cutoff for  $F(4, 968)$  at level 0.05 is actually 2.38; and the cutoff at level 0.01 would be 3.34.

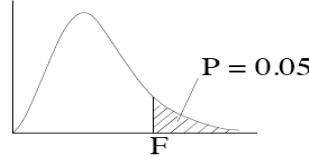
Table 14.5: ANOVA table for breastfeeding data: Full Scale IQ, Adjusted.

	SS	d.f.	MS	F
Between Samples	3597 (A)	4 (B)	894.8 ( $X = A/B$ )	3.81
Errors (Within Samples)	227000 (C)	968 (D)	234.6 ( $Y = C/D$ )	
Total	230600 (TSS=A+C)	972 ( $N - 1$ )		

#### 14.4.3 Another Example: Exercising rats

We consider the following example, adapted from [MM98, Chapter 15]. A study was performed to study the effect of exercise on bone density in rats. 30 rats were divided into three groups of ten: The first group carried out ten “high jumps” (60 cm) a day for eight weeks; the second group carried out ten “low jumps” (30 cm) a day for eight weeks; the third group had no special exercise. At the end of the treatment period, each rat’s bone density was measured. The results are given in Table 14.6.

We wish to test the null hypothesis that the different groups have the same mean bone density, against the alternative hypothesis that they have different bone densities. We first carry out the ANOVA analysis. The total sum of squares is 20013.4. The error sum of squares (ESS) is computed as  $9s_1^2 + 9s_2^2 + 9s_3^2 = 12579.5$ . The between-groups sum of squares (BSS) is computed as  $10(638.7 - 617.4)^2 + 10(612.5 - 617.4)^2 + 10(601.1 - 617.4)^2 = 7433.9$  (here the overall mean is 617.4). Note that indeed  $TSS = ESS + BSS$ . We complete the computations in Table 14.7, obtaining  $F = 7.98$ . Looking in the column for 2, and the row for 30 (since there is no row on your

Variance ratio  $F = s_1^2/s_2^2$  with  $\nu_1$  and  $\nu_2$  degrees of freedom respectively.

$\nu_2$	1	2	3	4	5	6	8	12	24	$\infty$	$\nu_2$
$\nu_1$	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67	6
6	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93	8
8	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54	10
10	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30	12
12	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13	14
14	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01	16
16	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92	18
18	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84	20
20	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62	30
30	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51	40
40	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.39	60
$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00	$\infty$

Figure 14.2: Table of F distribution; finding the cutoff at level 0.05 for the breastfeeding study.

table for 27), we see that the cutoff at level 0.05 is 3.32. Thus, we conclude that the difference in means between the groups is statistically significant.

## 14.5 Multifactor ANOVA

The procedure described in section 14.4 leads to obvious extensions. We have observations

$$x_{ki} = \mu_k + \epsilon_{ki}, \quad k = 1, \dots, K; \quad i = 1, \dots, n_k$$

where the  $\epsilon_{ki}$  are the normally distributed “errors”, and  $\mu_k$  is the true mean for group  $k$ . Thus, in the example of section 14.4.3, there were three groups, corresponding to three different exercise regimens, and ten different samples for each regimen. The obvious estimate for  $\mu_k$  is

$$\bar{x}_{k\cdot} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ki},$$

and we use the F test to determine whether the differences among the means are genuine. We decompose the total variance of the observations into the

(a) Full data											
	High	626	650	622	674	626	643	622	650	643	631
	Low	594	599	635	605	632	588	596	631	607	638
	Control	614	569	653	593	611	600	603	593	621	554

(b) Summary statistics			
	Group	Mean	SD
	High	638.7	16.6
	Low	612.5	19.3
	Control	601.1	27.4

Table 14.6: Bone density of rats after given exercise regime, in mg / cm<sup>3</sup>

portion that is between groups and the portion that is within groups. If the between-group variance is too big, we reject the hypothesis of equal means.

Many experiments naturally lend themselves to a two-way layout. For instance, there may be three different exercise regimens and two different diets. We represent the measurements as

$$x_{kji} = \mu_k + \nu_j + \epsilon_{kji}, \quad k = 1, 2, 3; \quad j = 1, 2; \quad i = 1, \dots, n_{kj}.$$

It is then slightly more complicated to isolate the exercise effect  $\mu_k$  and the diet effect  $\nu_j$ . We test for equality of these effects by again splitting the variance into pieces: the total sum of squares falls naturally into four pieces, corresponding to the variance over diets, variance over exercise regimens, variance over joint diet and exercise, and the remaining variance within each group. We then test for whether the ratios of these pieces are too far from the ratio of the degrees of freedom, as determined by the F distribution.

Multifactor ANOVA is quite common in experimental practice, but will not be covered in this course.

## 14.6 Kruskal-Wallis Test

Just as there is the non-parametric rank-sum test, similar to the t and z tests for equality of means, there is a non-parametric version of the F test, called the Kruskal-Wallis test. As with the rank-sum test, the basic idea is

Table 14.7: ANOVA table for rat exercise data.

	SS	d.f.	MS	F
Between Samples	7434 (A)	2 (B)	3717 ( $X = A/B$ )	7.98
Errors (Within Samples)	12580 (C)	27 (D)	466 ( $Y = C/D$ )	
Total	20014 (TSS=A+C)	29 ( $N - 1$ )		

simply to substitute ranks for the actual observed values. This avoids the assumption that the data were drawn from a normal distribution.

In Table 14.8 we duplicate the data from Table 14.6, replacing the measurements by the numbers 1 through 30, representing the ranks of the data: the lowest measurement is number 1, and the highest is number 30. In other words, suppose we have observed  $K$  different groups, with  $n_i$  observations in each group. We order all the observations in one large sequence of length  $N$ , from lowest to highest, and assign to each one its rank. (In case of ties, we assign the average rank.) We then sum the ranks in group  $i$ , obtaining numbers  $R_1, \dots, R_K$ . Then

The Kruskal-Wallis test statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^K \frac{R_i^2}{n_i} - 3(N+1).$$

Under the null hypothesis, that all the samples came from the same distribution,  $H$  has the  $\chi^2$  distribution with  $K - 1$  degrees of freedom.

In the rat exercise example, we have the values of  $R_i$  given in Table 14.7(b), yielding  $H = 10.7$ . If we are testing at the 0.05 significance level, the cutoff for  $\chi^2$  with 2 degrees of freedom is 5.99. Thus, we conclude again that there is a statistically significant difference among the distributions of bone density in the three groups.

(a) Full data											
High	18.5	27.5	16.5	30	18.5	25.5	16.5	27.5	25.5	20.5	
Low	6	8	23	11	22	3	7	20.5	12	24	
Control	14	2	29	4.5	13	9	10	4.5	15	1	

(b) Summary statistics	
Group	Sum
High	226.5
Low	136.5
Control	102

Table 14.8: Ranks of data in Table 14.6.



# Lecture 15

## Regression and correlation: Detecting trends

### 15.1 Introduction: Linear relationships between variables

In lectures 11 and 14 we have seen several examples of statistical problems which could be stated in this form: We have pairs of observations  $(x_i, y_i)$ , where  $y_i$  is numerical, and  $x_i$  categorical — that is,  $x_i$  is an assignment to one of a few possible categories — and we seek to establish the extent to which the distribution of  $y_i$  depends on the category  $x_i$ . For example, in section 11.1,  $y_i$  is a height and  $x_i$  is either “early-married” or “late-married”. In section 14.4.3 the  $y_i$  are measures of bone density, and the  $x_i$  are the experimental category (control, high exercise, low exercise). In section 14.4.2 the  $y_i$  are IQ measures, and the  $x_i$  are categories of length of breastfeeding.

The last example in particular points up the limitation of the approach we have taken so far. If we think that breastfeeding affects IQ, we would like to know if there is a linear relationship, and if so, how strong it is. That is, the data would be telling a very different story if group 5 infants ( $> 9$  months) wound up with the highest IQ and group 1 ( $\leq 1$  month) with the lowest; than if the high IQs were in groups 1, 2 and 4, with lower IQs in groups 3 and 5. But ANOVA only gives one answer: The group means are different.

The linear relationship is the simplest one we can test for, when the **covariate**  $x_i$  is numerical. Intuitively, it says that *each increase in  $x_i$  by one unit effects a change in  $y_i$  by the same fixed amount*. Formally, we write

$$\text{Regression Model } y_i = \beta x_i + \alpha + \epsilon_i.$$

We think of this as a set of observations from random variables  $(X, Y)$  that satisfy the relationship  $Y = \beta X + \alpha + E$ , where  $E$  is independent of  $X$ . We call this a **linear relation** because the pairs of points  $(x_i, y_i)$  with  $y_i = \beta x_i + \alpha$  lie approximately on a straight line. Here the  $\epsilon_i$  are “noise” or “error” terms. They are the random variation in measurement of  $Y$  that prevent us from seeing the data lying exactly (and obviously) on a line  $y = \beta x + \alpha$ . They may represent actual random errors from a “true” value of  $Y$  that is exactly  $\beta X + \alpha$ , or it may mean that  $\beta X + \alpha$  is an overall trend, with  $E$  reflecting the contribution of other factors that have nothing to do with  $X$ . For instance, in the breastfeeding example, we are interested to see whether children who received more breastmilk had higher IQs — but we don’t expect children who were nursed for the same length of time to end up all with the same IQ, as they will differ genetically, socially, and in simple random proclivities. (As mentioned in section 14.4.2, the change from raw to adjusted mean IQ scores is intended to compensate for some of the more systematic contributions to the “error term”: mothers’ smoking, parents’ IQ and income, and so forth. The effect is to reduce the size of the  $\epsilon_i$  terms, and so (one hopes) to make the true trend  $\beta$  more apparent.) Of course, if the  $\epsilon_i$  are large on average — if  $E$  has a high variance relative to the variance of  $X$  — then the linear relationship will be drowned in a sea of noise.

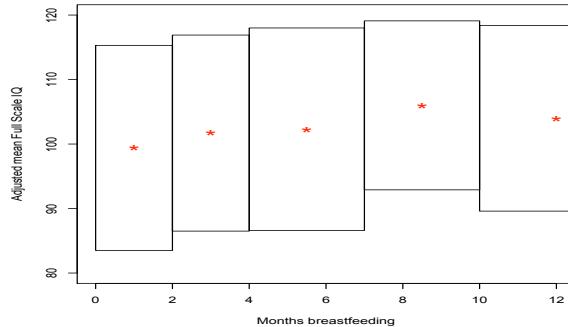


Figure 15.1: Plot of data from the breastfeeding IQ study in Table 14.1. Stars represent mean for the class, boxes represent mean  $\pm 2$  Standard Errors.

## 15.2 Scatterplots

The most immediate thing that we may wish to do is to get a picture of the data with a **scatterplot**. Some examples are shown in Figure 15.2.

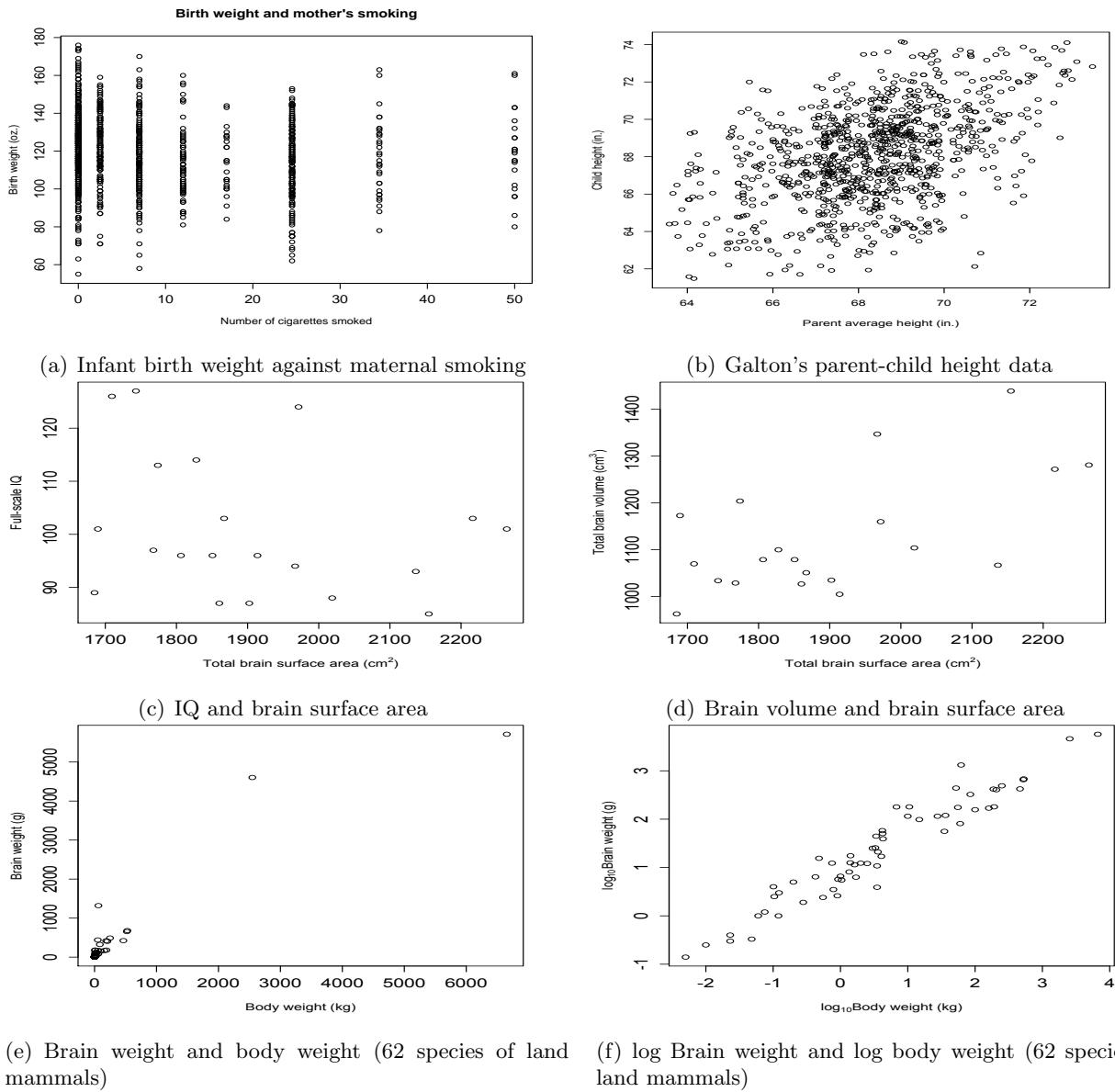


Figure 15.2: Examples of scatterplots

### 15.3 Correlation: Definition and interpretation

Given two paired random variables  $(X, Y)$ , with means  $\bar{X}$  and  $\bar{Y}$ , we define the **covariance** of  $X$  and  $Y$ , to be

$$\text{Cov}(X, Y) = \text{mean}[(X - \bar{X})(Y - \bar{Y})].$$

For  $n$  paired observations  $(x_i, y_i)$ , with means  $\bar{x}$  and  $\bar{y}$ , we define the **population covariance**

$$c_{xy} = \text{mean}[(x_i - \bar{x})(y_i - \bar{y})] = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

As with the SD, we usually work with the **sample covariance**, which is just

$$s_{xy} = \frac{n}{n-1} c_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

This is a better estimate for the covariance of the random variables that  $x_i$  and  $y_i$  are sampled from.

Notice that the means of  $x_i - \bar{x}$  and  $y_i - \bar{y}$  are both 0: On average,  $x_i$  is neither higher nor lower than  $\bar{x}$ . Why is the covariance then not also 0? If  $X$  and  $Y$  are independent, then each value of  $X$  will come with, on average, the same distribution of  $Y$ 's, so the positives and negatives will cancel out, and the covariance will indeed be 0. On the other hand, if high values of  $x_i$  tend to come with high values of  $y_i$ , and low values with low values, then the product  $(x_i - \bar{x})(y_i - \bar{y})$  will tend to be positive, making the covariance positive.

While positive and negative covariance have obvious interpretations, the magnitude of covariance does not say anything straightforward about the strength of connection between the covariates. After all, if we simply measure heights in millimetres rather than centimetres, all the numbers will become 10 times as big, and the covariance will be multiplied by 100. For this reason, we normalise the covariance by dividing it by the product of the two standard deviations, producing the quantity called **correlation**:

$$\text{Correlation}$$

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Of course, we estimate the correlation in a corresponding way:

Sample correlation

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

It is easy to see that correlation does not change when we rescale the data — for instance, by changing the unit of measurement. If  $x_i$  were universally replaced by  $x_i^* = \alpha x_i + \beta$ , then  $s_{xy}$  becomes  $s_{xy}^* = \alpha s_{xy}$ , and  $s_x$  becomes  $s_x^* = \alpha s_x$ . Since the extra factor of  $\alpha$  appears in the numerator and in the denominator, the final result of  $r_{xy}$  remains unchanged. In fact, it turns out that  $r_{xy}$  is always between  $-1$  and  $1$ . The correlation of  $-1$  means that there is a perfect linear relationship between  $x$  and  $y$  with negative sign; correlation of  $+1$  means that there is a perfect linear relationship between  $x$  and  $y$  with positive sign; and correlation  $0$  means no linear relationship at all.

In Figure 15.3 we show some samples of standard normally distributed pairs of random variables with different correlations. As you can see, high positive correlation means the points lie close to an upward-sloping line; high negative correlation means the points lie close to a downward-sloping line; and correlation close to  $0$  means the points lie scattered about a disk.

## 15.4 Computing correlation

There are several alternative formulae for the covariance, which may be more convenient than the standard formula:

$$\begin{aligned} s_{xy} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \frac{n}{n-1} \bar{x} \bar{y} \\ &= \frac{1}{4} (s_{x+y}^2 - s_{x-y}^2), \end{aligned}$$

where  $s_{x+y}$  and  $s_{x-y}$  are the sample SDs of the collections  $(x_i + y_i)$  and  $(x_i - y_i)$  respectively.

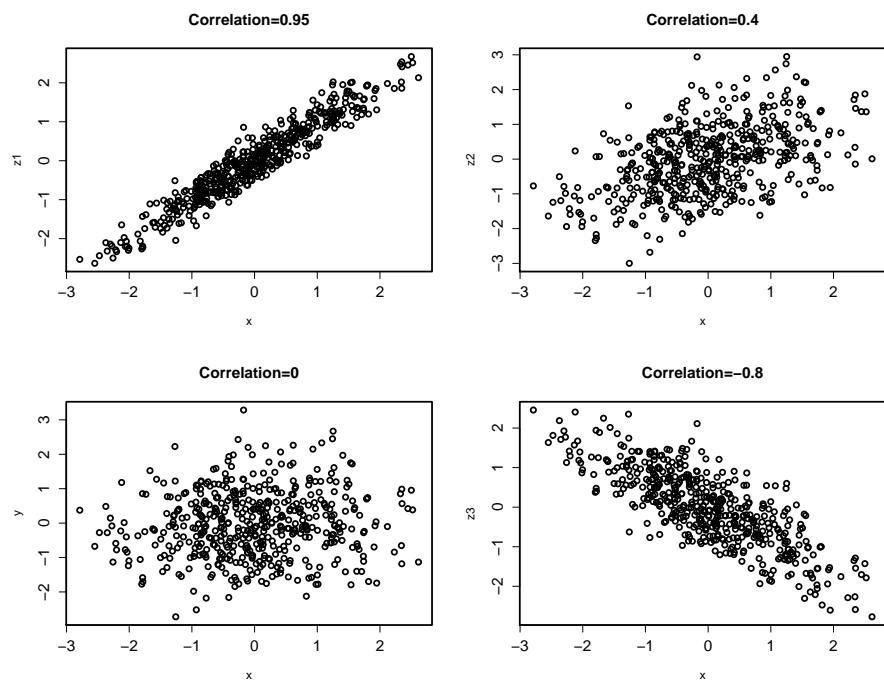


Figure 15.3: Examples of pairs of random variables with different correlations.

### 15.4.1 Brain measurements and IQ

A study was done [TLG<sup>+</sup>98] to compare various brain measurements and IQ scores among 20 subjects.<sup>1</sup> In Table 15.1 we give some of the data, including a measure of full-scale IQ and MRI estimates of total brain volume and total brain surface area.

Table 15.1: Brain measurement data.

Subject	Brain volume (cm <sup>3</sup> )	Brain surface area (cm <sup>2</sup> )	IQ
1	1005	1914	96
2	963	1685	89
3	1035	1902	87
4	1027	1860	87
5	1281	2264	101
6	1272	2216	103
7	1051	1867	103
8	1079	1851	96
9	1034	1743	127
10	1070	1709	126
11	1173	1690	101
12	1079	1806	96
13	1067	2136	93
14	1104	2019	88
15	1347	1967	94
16	1439	2155	85
17	1029	1768	97
18	1100	1828	114
19	1204	1774	113
20	1160	1972	124
Mean	1125	1906	101
SD	175	125	13.2

In Figures 15.2(c) and 15.2(d) we see scatterplots of IQ against surface

<sup>1</sup>In fact, the 20 subjects comprised 5 pairs of male and 5 pairs of female monozygous twins, so there are plenty of interesting possibilities for factorial analysis. The data are available in a convenient table at [http://lib.stat.cmu.edu/datasets/IQ\\_Brain\\_Size](http://lib.stat.cmu.edu/datasets/IQ_Brain_Size).

area and volume against surface area, respectively. There seems to be a negative relationship between  $IQ$  and surface area, and a positive relationship between volume and surface area — the latter, while not surprising, being hardly inevitable. Given the triples of numbers it is straightforward to compute all three paired correlations. If we denote the volume, surface area, and IQ by  $x_i$ ,  $y_i$ , and  $z_i$ , we can compute

$$s_{xy} = 13130, \quad s_{yz} = -673, \quad s_{xz} = -105.$$

This leads us to

$$r_{xy} = \frac{13130}{s_x s_y} = 0.601.$$

Similarly

$$r_{yz} = -0.291, \quad r_{xz} = -0.063.$$

### 15.4.2 Galton parent-child data

There is a famous collection of data collected by Francis Galton, measuring heights and lengths of various body parts for various related individuals. (This was one of the earliest quantitative studies of human inheritance.) We consider the heights of parents and children.<sup>2</sup> Parent height is given as the average height of the two parents.<sup>3</sup> The data were given in units of whole inches, so we have “jittered” the data to make the scatterplot of Figure 15.2(b). That is, we have added some random noise to each of the values when it was plotted, so that the dots don’t lie right on top of each other.

Suppose we have the list of parent height and child heights; and the list of differences between parent and child heights and sum of parent and child heights. We know the variances for each of these: Then we have

$$s_{xy} = \frac{1}{4}(13.66 - 5.41) = 2.07,$$

and

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{2.07}{2.52 \cdot 1.79} = 0.459.$$

---

<sup>2</sup>Available as the dataset `galton` of the `UsingR` package of the R programming language, or directly from <http://www.bun.kyoto-u.ac.jp/~suchii/galton86.html>.

<sup>3</sup>We do not need to pay attention to the fact that Galton multiplied all female heights by 1.08.

Table 15.2: Variances for different combinations of the Galton height data.

	SD	Variance
Parent	1.79	3.19
Child	2.52	6.34
Sum	3.70	13.66
Difference	2.33	5.41

### 15.4.3 Breastfeeding example

This example is somewhat more involved than the others, and than the kinds of covariance computations that appear on exams.

Consider the breastfeeding–IQ data (Table 14.1), which we summarise in the top rows of Table 15.3. We don’t have the original data, but we can use the above formulas to estimate the covariance, and hence the correlation, between number of months breastfeeding and adult IQ. Here  $x_i$  is the number of months individual  $i$  was breastfed, and  $y_i$  the adult IQ.

Table 15.3: Computing breastfeeding–Full-IQ covariance for Copenhagen infant study.

Average number of months breastfeeding					
	1	3	5.5	8.5	> 9
N	272	305	269	104	23
SD	15.9	15.2	15.7	13.1	14.4
Adjusted Mean	99.4	101.7	102.3	106.0	104.0
Contribution to $\sum x_i y_i$	27037	93056	151353	93704	28704

In the bottom row of the table we give the contribution that those individuals made to the total  $\sum x_i y_i$ . Since the  $x_i$  are all about the same in any column, we treat them as though they were all about the same, equal to the average  $x$  value in the column. (We give these averages in the first row. This involves a certain amount of guesswork, particularly for the last column; on the other hand, there are very few individuals in that column.)

The sum of the  $y$  values in any column is the average  $y$  value multiplied by the relevant number of samples. Consider the first column:

$$\begin{aligned}\sum_{i \text{ in first column}} x_i y_i &\approx 1 \cdot \sum_{i \text{ in first column}} y_i \\&= 1 \cdot N_1 \bar{y}_1 \\&= 1 \cdot 272 \cdot 99.4 \\&= 27037.\end{aligned}$$

Similarly, the second column contributes  $3 \cdot 305 \cdot 101.7 = 93065$ , and so on. Adding these contributions yields  $\sum x_i y_i = 393853$ .

We next estimate  $\bar{x}$ , treating it as though there were 272  $x_i = 1$ , 305  $x_i = 3$ , and so on, yielding

$$\bar{x} = \frac{1}{973} (272 \cdot 1 + 305 \cdot 3 + 269 \cdot 5.5 + 104 \cdot 8.5 + 27 \cdot 12) = 3.93.$$

To estimate  $\bar{y}$ , we take

$$\sum_i = \sum_{i \text{ in column 1}} y_i + \sum_{i \text{ in column 2}} y_i + \sum_{i \text{ in column 3}} y_i + \sum_{i \text{ in column 4}} y_i + \sum_{i \text{ in column 5}} y_i.$$

The sum in a column is just the average in the column multiplied by the number of observations in the column, so we get

$$\bar{y} = \frac{1}{973} (272 \cdot 99.4 + 305 \cdot 101.7 + 269 \cdot 102.3 + 104 \cdot 106.0 + 27 \cdot 104.0) = 101.74.$$

Thus, we get

$$s_{xy} = \frac{393853}{972} - \frac{973}{972} \cdot 101.74 \cdot 3.93 = 4.95$$

To compute the correlation we now need to estimate the variance of  $x$  separately. (The variance of  $y$  could be computed from the other given data — using the individual column variances to compute  $\sum y_i^2$  — but we are given that  $s_y = 15.7$ .) For the  $x$  values, we continue to treat them as though they were all the same in a column, and use the formula for grouped data. We get then

$$\begin{aligned}s_x^2 &= \frac{1}{972} \left( 272 \cdot (1 - 3.93)^2 + 305 \cdot (3 - 3.93)^2 + 269 \cdot (5.5 - 3.93)^2 \right. \\&\quad \left. + 104 \cdot (8.5 - 3.93)^2 + 27 \cdot (12 - 3.93)^2 \right) \\&= 7.13.\end{aligned}$$

Thus, we have

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{4.95}{2.67 \cdot 15.7} = 0.118.$$

## 15.5 Testing correlation

One thing we want to do is to test whether the correlation between two variables is different from 0, or whether the apparent correlation could be merely due to chance variation. After all, if we sample  $X$  and  $Y$  at random, completely independently, the correlation won't come out to be *exactly* 0.

We suppose we have a sample  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , from random variables  $(X, Y)$ , which we assume are normally distributed, but with unknown means, variances, and correlation. We formulate the null hypothesis

$$H_0 : \rho_{XY} = 0,$$

and test it against the alternative hypothesis that  $\rho_{XY} \neq 0$ . As usual, we need to find a test statistic  $R$  that has two properties

- (1). Extreme values of  $R$  correspond to extreme failure of the null hypothesis (relative to the alternative). In other words,  $R$  should tend to take on more extreme values when  $\rho_{XY}$  is farther away from 0.
- (2). We know the distribution of  $R$  (at least approximately).

In this case, our test statistic is

$$R = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}.$$

It can be shown that, under the null hypothesis, this  $R$  has the Student  $t$  distribution with  $n - 2$  degrees of freedom. Thus, we can look up the appropriate critical value, and reject the null hypothesis if  $|R|$  is above this cutoff. For example, in the Brain measurement experiments of section 15.4.1 we have correlation between brain volume and surface area being 0.601 from 20 samples, which produces  $R = 3.18$ , well above the threshold value for  $t$  with 18 degrees of freedom at the 0.05 level, which is 2.10. On the other hand, the correlation  $-0.291$  for surface area against IQ yields  $R = 1.29$ , which does not allow us to reject the null hypothesis that the true underlying population correlation is 0; and the correlation  $-0.063$  between volume and IQ yields  $R$  only 0.255.

Note that for a given  $n$  and choice of level, the threshold in  $t$  translates directly to a threshold in  $r$ . If  $t_*$  is the appropriate threshold value in  $t$ , then we reject the null hypothesis when our sample correlation  $r$  is larger than  $\sqrt{t_*/(n - 2 + t_*)}$ . In particular, for large  $n$  and  $\alpha = 0.05$ , we have a threshold for  $t$  very close to 2, so that we reject the null hypothesis when  $|r| > \sqrt{2/n}$ .

## 15.6 The regression line

### 15.6.1 The SD line

One way of understanding the relationship between  $X$  and  $Y$  is to ask, if you know  $X$ , how much does that help you to predict the corresponding value of  $Y$ ? For instance, in section 15.4.2 we might want to know what the predicted adult height should be for a child, given the height of the parents. Presumably, average-height parents (68.3 inches) should have average-height children (68.1 inches). But what about parents who average 70 inches (remember, the “parent height” is an average of the mother and father): they are a bit taller than average, so we expect them to have children who are taller than average, but by how much?

The most naïve approach would be to say that parents who are 1.7 inches above the mean should have children who are 1.7 inches above the mean. This can’t be right, though, because there is considerably more spread in the children’s heights than in the parents’ heights<sup>4</sup> — the parents’ SD is 1.79, where the children’s SD is 2.50.

One approach would be to say, the parents are about 1 SD above the mean for parents (actually, 0.95 SD), so the children should be about 1 SD above the mean for children. If we make that prediction for all parents, and plot the corresponding prediction for their children, we get the green line in Figure 15.4. We follow [FPP98] in calling this the **SD line**. This is a line that passes through the point corresponding to the means of the two variables, and rises one SD in  $Y$  for every SD in  $X$ . If we think of the cloud of points as an oval, the SD line runs down the long axis of the oval. The formula for the SD line is then

$$Y - \bar{y} = \frac{s_y}{s_x}(X - \bar{x}). \quad (15.1)$$

---

<sup>4</sup>This is because the parent height is the average of two individuals. We note that the parents’ SD is almost exactly  $1/\sqrt{2}$  times the children’s SD.

This can be rearranged to

$$Y = \frac{s_y}{s_x}X + \left( \bar{y} - \frac{s_y}{s_x}\bar{x} \right).$$

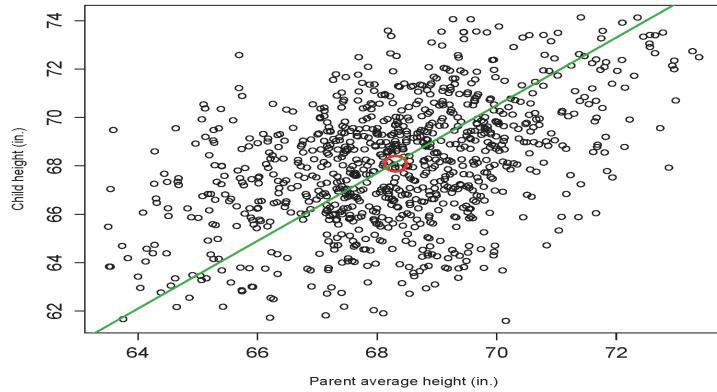


Figure 15.4: Galton parent-child heights, with SD line in green. The point of the means is shown as a red circle.

### 15.6.2 The regression line

On further reflection, though, it becomes clear that the SD line can't be the best prediction of  $Y$  from  $X$ . If we look at Figure 15.3, we see that the line running down the middle of the cloud of points is a good predictor of  $Y$  from  $X$  if the correlation is close to 1. When the correlation is 0, though, we'd be better off ignoring  $X$  and predicting  $Y$  to be  $\bar{y}$  always; and, of course, when the correlation is negative, the line really needs to slope in the other direction.

What about intermediate cases, like the Galton data, where the correlation is 0.46? One way of understanding this is to look at a narrow range of  $X$  values (parents' heights), and consider what the corresponding range of  $Y$  values is. In figure 15.5, we sketch in rectangles showing the approximate range of  $Y$  values corresponding to  $X = 66, 68, 70$ , and  $72$  inches. As you can see the middle of the  $X = 66$  range is substantially above the SD line, whereas the middle of the  $X = 72$  range is below the SD line. This makes sense, since the connection between the parents'

and childrens' height is not perfect. If the linear relation  $Y = \beta X + \alpha$  held exactly, then  $SD(Y) = \beta SD(X)$ , and if  $X = \bar{X} + SD(X)$ , then  $Y = \bar{Y} + \beta SD(X) = \bar{Y} + SD(Y)$  exactly. But when the  $Y$  values are spread out around  $\beta X + \alpha$ , the more  $SD(Y)$  gets inflated by irrelevant noise that has nothing to do with  $X$ . This means that an increase of 1 SD in  $X$  won't contribute a full SD to  $Y$ , on average, but something less than that. The line that runs approximately through the midpoints of the columns, shown in blue in Figure 15.5, is called the **regression line**. So how many SD's of  $Y$  is a 1 SD change in  $X$  worth? It turns out, this is exactly the correlation.

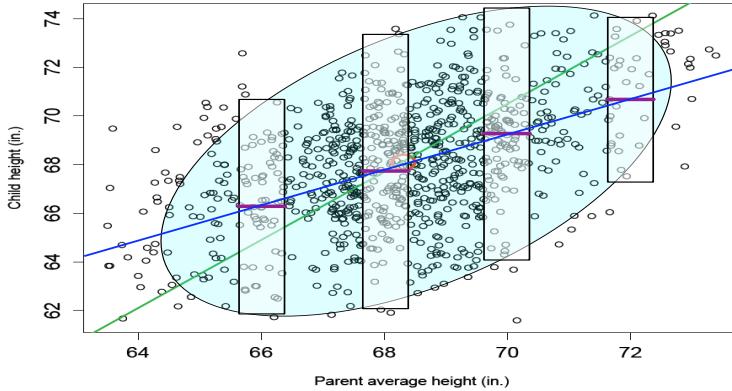


Figure 15.5: The parent-child heights, with an oval representing the general range of values in the scatterplot. The SD line is green, the regression line blue, and the rectangles represent the approximate span of  $Y$  values corresponding to  $X = 66, 68, 70, 72$  inches.

The formula for the regression line is

$$(Y - \bar{y}) = r_{xy} \frac{s_y}{s_x} (X - \bar{x}) = \frac{s_{xy}}{s_x^2} (X - \bar{x}).$$

This is equivalent to

$$Y = bX + a, \text{ where } b = \frac{s_{xy}}{s_x^2} \text{ and } a = \left( \bar{y} - \frac{s_y}{s_x} \bar{x} \right).$$

Another way of understanding the regression line is as the answer to the question: Suppose we want to predict  $y_i$  from  $x_i$  with a linear relation  $y = bx + a$ . Which choice of  $a$  and  $b$  will make the total squared error as small as possible. That is, the regression line makes  $\sum(y_i - (bx_i + a))^2$  as small as possible. Note that the choice of the total square error as the way to “score” the prediction errors implies a certain choice about how much we care about a few big errors relative to a lot of small errors. The regression line tries, up to a point, to make the biggest error as small as possible, at the expense of making some of the smaller errors bigger than they might have been.

### Example 15.1: Regression line(s) for Galton’s data

For Galton’s parent-child height data, the correlation was 0.46, so that the slope of the regression line is

$$b = \frac{s_y}{s_x} r_{xy} = 0.649,$$

$$a = \bar{y} - b\bar{x} = 23.8.$$

Thus, if we have parents with average height 72 inches, we would predict that their child should have, on average, height

$$y_{pred} = 72 \times 0.649 + 23.8 = 70.5 \text{ inches.}$$

Suppose we reverse the question: Given a child of height 70.5 inches, what should we predict for the average of his or her parents’ heights? You might suppose it is 72 inches, simply reversing the previous calculation. In fact, though, we need to redo the calculation. All predictions must be closer to the (appropriate) mean than the observations of the independent variable on which the prediction is based. The child is  $(70.5 - 68.1)/2.52 = 0.95$  SDs away from the mean, and the prediction for the parents must be closer to the parents’ mean than that. In fact, if we write the coefficients for the regression line predicting *parents from children* as  $b'$  and  $a'$ , we have

$$b' = \frac{s_{xy}}{s_x s_y} = 0.326,$$

$$a' = \bar{x} - b'\bar{y} = 45.9.$$

Thus, the prediction for the parents' heights is  $45.9 - 0.326 \times 70.5 = 68.9$  inches. ■

This may be seen in the following simple example. Suppose we have five paired observations:

Table 15.4: Simple regression example

					mean	SD
$x$	6	2	5	7	4.8	1.9
$y$	3	2	4	6	4.0	1.58
prediction						
$0.54x + 1.41$	4.65	2.49	4.11	5.19	3.57	
residual	-1.65	-0.49	-0.11	0.81	1.43	

We compute the sample covariance as

$$s_{xy} = \frac{1}{5-1} \left( \sum x_i y_i - 5\bar{x}\bar{y} \right) = \frac{1}{y} (104 - 5 \cdot 4.8 \cdot 4.0) = 2.0,$$

yielding

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0.66.$$

The regression line then has coefficients

$$b = \frac{s_{xy}}{s_x^2} = 0.54, \text{ and } a = \bar{y} - b\bar{x} = 1.41.$$

This is the line plotted in blue in Figure 15.6; in green is the SD line. We have shown with dashed lines the size of the errors that would accrue if the one or the other line were used for predicting  $y$  from  $x$ . Note that some of the errors are smaller for the SD line, but the largest errors are made larger still, which is why the regression line has a smaller total squared error. In Figure 15.8 we do the same thing for the brain volumes and surface areas of Table 15.1.

### 15.6.3 Confidence interval for the slope

Suppose the observations  $(x_i, y_i)$  are independent observations of normal random variables  $(X, Y)$ , which satisfy

$$Y = \beta X + \alpha + \epsilon,$$

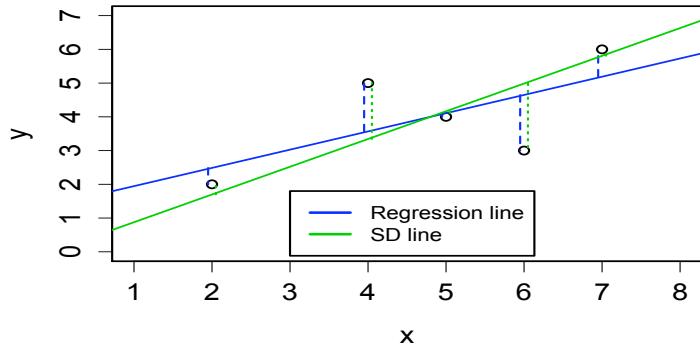


Figure 15.6: A scatterplot of the hypothetical data from Table 15.4. The regression line is shown in blue, the SD line in green. The dashed lines show the prediction errors for each data point corresponding to the two lines.

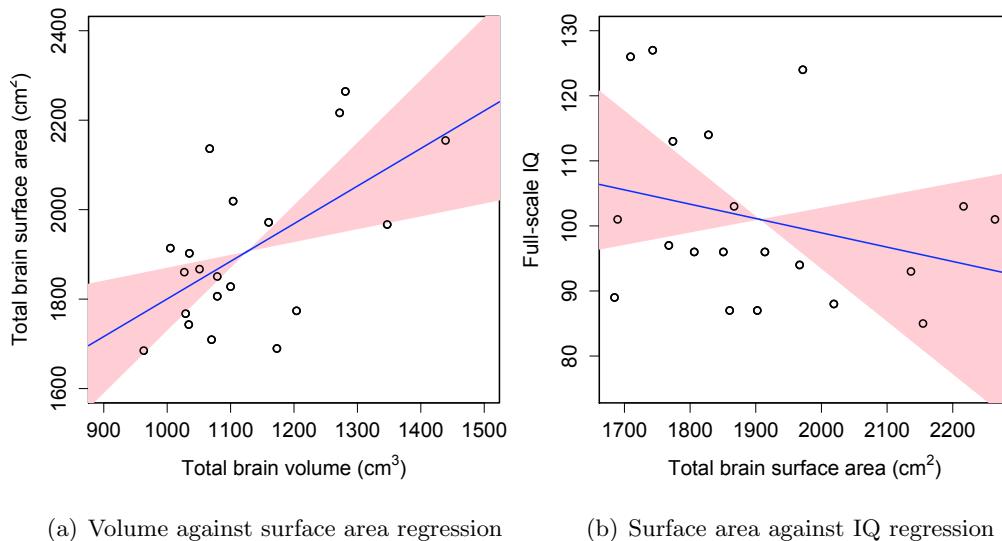


Figure 15.7: Regression lines for predicting surface area from volume and predicting IQ from surface area. Pink shaded region shows confidence interval for slope of the regression line.

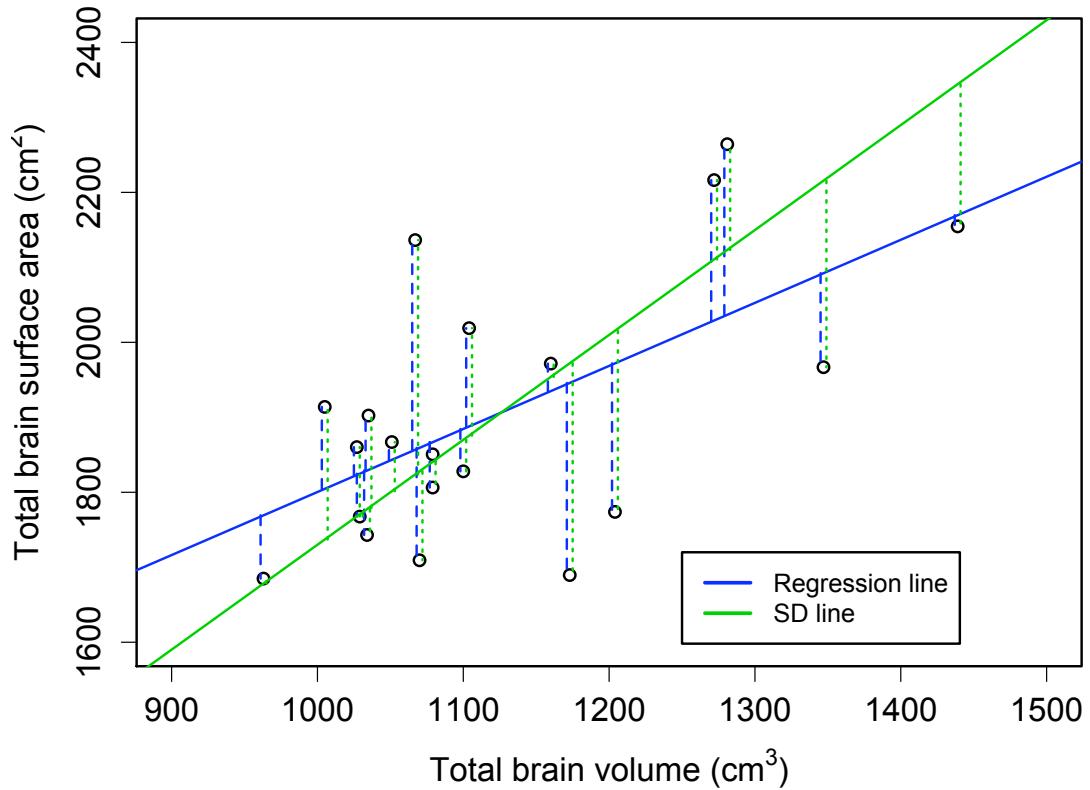


Figure 15.8: A scatterplot of brain surface area against brain volume, from the data of Table 15.1. The regression line is shown in blue, the SD line in green. The dashed lines show the prediction errors for each data point corresponding to the two lines.

where  $\epsilon$  is a normal “error term” independent of  $X$ . The idea is that  $Y$  gets a contribution from  $X$ , and then there is the contribution from  $\epsilon$ , representing “everything else” that has nothing to do with  $X$ . The regression coefficients  $b$  and  $a$  will be estimates for the true slope and intercept  $\beta$  and  $\alpha$ . We then ask the standard question: How certain are we of these estimates?

One thing we may want to do is to test whether there is really a nonzero slope, or whether the apparent difference from  $\beta = 0$  might be purely due to chance variation. Since the slope is zero exactly when the correlation is zero, this turns out to be exactly the same as the hypothesis test for zero correlation, as described in section 15.5.

The same ideas can be used to produce a confidence interval for  $\beta$ . Under the normality assumption, it can be shown that the random estimate  $b$  has standard error

$$SE(b) \approx \frac{b\sqrt{1-r^2}}{r\sqrt{n-2}},$$

and that

$$t := \frac{\beta - b}{SE(b)}$$

has approximately the  $t$  distribution with  $n - 2$  degrees of freedom. Thus, a  $(1 - \alpha) \times 100\%$  confidence interval for  $\beta$  is

$$b \pm T \cdot SE(b),$$

where  $T$  is chosen from the  $t$  distribution table to be the threshold value for the test at level  $\alpha$  with  $n - 2$  degrees of freedom; in other words,  $T = t_{1-\alpha/2}(n - 2)$ , the value such that the  $t$  distribution has a total of  $\alpha/2$  probability above  $T$  (and another  $\alpha/2$  probability below  $-T$ ).

#### 15.6.4 Example: Brain measurements

Suppose we know an individual’s brain volume  $x$ , and wish to make a best guess about that individual’s brain surface area. We have already computed the correlation to be 0.601 in section 15.4.1. Combining this with the means and SDs from Table 15.1, we get the regression coefficients

$$b = r_{xy} \frac{s_y}{s_x} = 0.601 \cdot \frac{175}{125} = 0.841 \quad a = \bar{y} - b\bar{x} = 959$$

The standard error for  $b$  is

$$SE(b) \approx \frac{b\sqrt{1-r^2}}{r\sqrt{n-2}} = \frac{0.841 \cdot \sqrt{1-0.601^2}}{0.601\sqrt{18}} = 0.264.$$

We look on the t-distribution table in the row for 18 degrees of freedom and the column for  $p = 0.05$  (see Figure 15.9) we see that 95% of the probability lies between  $-2.10$  and  $+2.10$ , so that a 95% confidence interval for  $\beta$  is

$$b \pm 2.10 \cdot SE(b) = 0.841 \pm 2.10 \cdot 0.264 = (0.287, 1.40).$$

The scatterplot is shown with the regression line  $y = .841x + 959$  in Figure 15.7(a), and the range of slopes corresponding to the 95% confidence interval is shown by the pink shaded region. (Of course, to really understand the uncertainty of the estimates, we would have to consider simultaneously the random error in estimating the means, hence the intercept of the line. This leads to the concept of a two-dimensional *confidence region*, which is beyond the scope of this course.)

Similarly, for predicting IQ from surface area we have

$$b = r_{yz} \frac{s_z}{s_y} = -0.291 \frac{13.2}{125} = -0.031, \quad a = \bar{z} - b\bar{y} = 160.$$

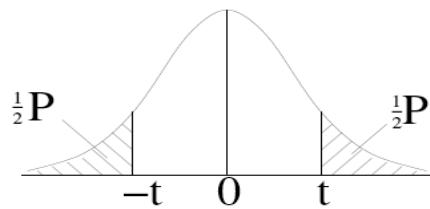
The standard error for  $b$  is

$$SE(b) \approx \frac{b\sqrt{1 - r^2}}{r\sqrt{n - 2}} = \frac{0.031 \cdot \sqrt{1 - 0.291^2}}{0.291\sqrt{18}} = 0.024.$$

A 95% confidence interval for the true slope  $\beta$  is then given by  $-0.031 \pm 2.10 \cdot 0.024 = (-0.081, 0.019)$ . The range of possible predictions of  $y$  from  $x$  — pretending, again, that we know the population means exactly — is given in Figure 15.7(b).

What this means is that each change of  $1 \text{ cm}^3$  in brain volume is typically associated with a change of  $0.841 \text{ cm}^2$  in brain surface area. A person of average brain volume —  $1126 \text{ cm}^3$  — would be expected to have average brain surface area —  $1906 \text{ cm}^2$  — but if we know that someone has brain volume  $1226 \text{ cm}^3$ , we would do well to guess that he has a brain surface area of  $1990 \text{ cm}^2$  ( $1990 = 1906 + 0.841 \cdot 100 = .841 \cdot 1226 + 959$ ). However, given sampling variation the number of  $\text{cm}^2$  typically associated with  $1 \text{ cm}^3$  change in volume might really be as low as  $0.287$  or as high as  $1.40$ , with 95% confidence.

Similarly, a person of average brain surface area  $1906 \text{ cm}^2$  might be predicted to have average IQ of 101, but someone whose brain surface area is found to be  $100 \text{ cm}^2$  might be predicted to have IQ below average by 3.1 points, so 97.9. At the same time, we can only be 95% certain that the change associated with  $100 \text{ cm}^2$  increase in brain surface area is between  $-8.1$  and  $+1.9$  points — hence, it might just as well be 0. We say that the *correlation between IQ and brain surface area is not statistically significant*, or that the slope of the regression line is not significantly different from 0.



Probability  $P$  of lying outside  $\pm t$

d.f.	P=0.10	P=0.05	P=0.02	P=0.01
1	6.31	12.71	31.82	63.7
2	2.92	4.30	6.96	9.93
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.90	2.37	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
11	1.80	2.20	2.72	3.11
12	1.78	2.18	2.68	3.06
13	1.77	2.16	2.65	3.01
14	1.76	2.15	2.62	2.98
15	1.75	2.13	2.60	2.95
16	1.75	2.12	2.58	2.92
17	1.74	2.11	2.57	2.90
18	1.73	2.10	2.55	2.88
19	1.73	2.09	2.54	2.86
20	1.73	2.09	2.53	2.85

Figure 15.9: T table for confidence intervals for slopes computed in section  
15.6.4



## Lecture 16

# Regression, Continued

### 16.1 $R^2$

What does it mean to say that  $bx_i + a$  is a good predictor of  $y_i$  from  $x_i$ ? One way of interpreting this would be to say that we will typically make smaller errors by using this predictor, than if we tried to predict  $y_i$  without taking account of the corresponding value of  $x_i$ .

Suppose we have our standard regression probability model  $Y = \beta X + \alpha + E$ : this means that the observations are

$$y_i = \beta x_i + \alpha + \epsilon_i.$$

Of course, we don't really get to observe the  $\epsilon$  terms: they are only inferred from the relationship between  $x_i$  and  $y_i$ . But if we have  $X$  independent of  $E$ , then we can use our rules for computing variance to see that

$$\text{Var}(Y) = \beta^2 \text{Var}(X) + \text{Var}(E). \quad (16.1)$$

If we think of variance as being a measure of uncertainty, then this says that the uncertainty about  $Y$  can be divided up into two parts: One part that comes from the uncertainty about  $X$ , and would be cleared up once we know  $X$ , and a **residual** uncertainty, that remains independent of  $X$ .<sup>1</sup> From our formula for the regression coefficients, we see that  $\beta^2 = r^2 \text{Var}(Y)/\text{Var}(X)$ . This means that the first term on the right in expression (16.1) becomes  $r^2 \text{Var}(Y)$ . In other words, the portion of the variance that is due to variability in  $X$  is  $r^2 \text{Var}(Y)$ , and the residual variance — the variance of the

---

<sup>1</sup>If this sounds a lot like the ANOVA approach from lecture 14, that's because it is. Formally, they're variants of the same thing, though developing this equivalence is beyond the scope of this course.

“error” term — is  $(1 - r^2)Var(Y)$ . Often this relation is summarised by the statement: “ $X$  **explains**  $r^2 \times 100\%$  of the variance.”

### 16.1.1 Example: Parent-Child heights

From the Galton data of section 15.4.2, we see that the variance of the child’s height is 6.34. Since  $r^2 = 0.21$ , we say that the parents’ heights **explain 21% of the variance in child’s height**. We expect the residual variance to be about  $6.34 \times 0.79 = 5.01$ . What this means is that the variance among children whose parents were all about the same height should be 5.01. In Figure 16.1 we see histograms of the heights of children whose parents all had heights in the same range of  $\pm 1$  inch. Not surprisingly, there is some variation in the shapes of these histograms, vary somewhat, but the variances are all substantially smaller than 6.34, a varying between 4.45 and 5.75.

### 16.1.2 Example: Breastfeeding and IQ

In section 15.4.3 we computed the correlation between number of months breastfeeding and adult IQ to be 0.118. This gives us  $r^2 = 0.014$ , so we say that the length of breastfeeding accounts for about 1.4% of adult IQ. The variance in IQ among children who were nursed for about the same length of time is about 1% less than the overall variance in IQ among the population. Not a very big effect, in other words. Is the effect real at all, or could it be an illusion, due to chance variation? We perform a hypothesis test, computing

$$R = \frac{r\sqrt{973 - 2}}{\sqrt{1 - r^2}} = 3.70.$$

The threshold for rejecting  $t$  with 971 degrees of freedom (the  $\infty$  row — essentially the same as the normal distribution) at the  $\alpha = 0.01$  significance level is 2.58. Hence, the correlation is *highly significant*. This is a good example of where “significant” in the statistical sense should not be confused with “important”. The difference is significant because the sample is so large that it is very unlikely that we would have seen such a correlation purely by chance if the true correlation were zero. On the other hand, explaining 1% of the variance is unlikely to be seen as a highly useful finding. (At the same time, it might be at least theoretically interesting to discover that there is any detectable effect at all.)

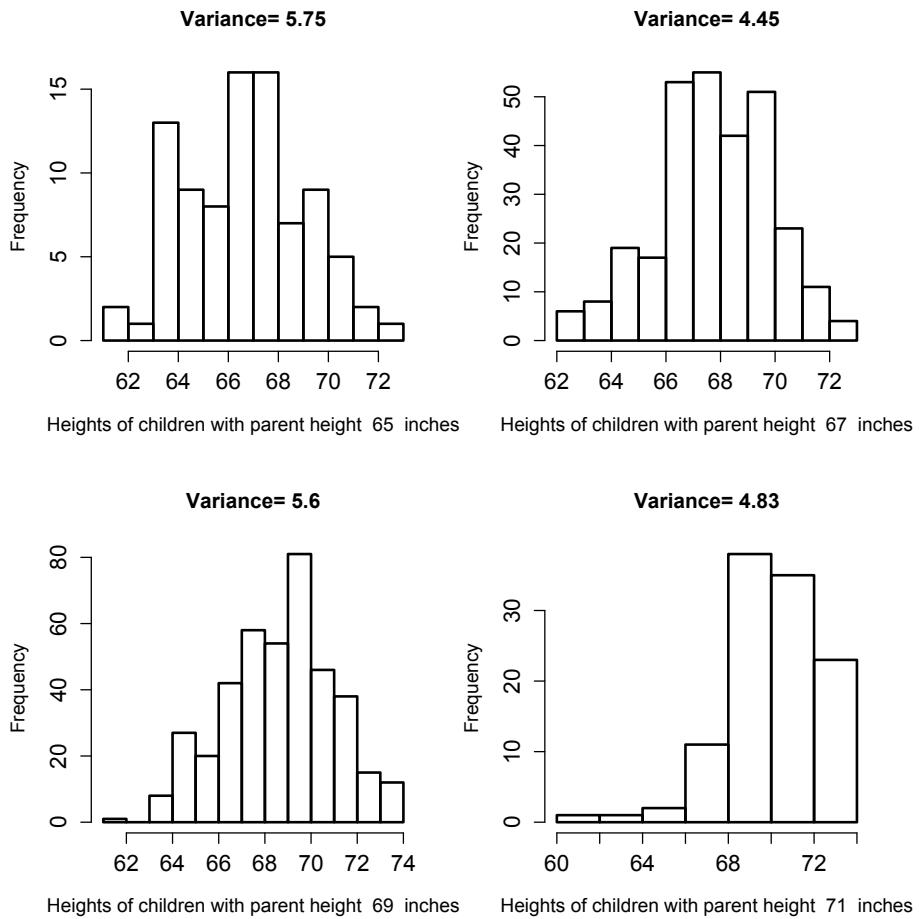


Figure 16.1: Histograms of Galton's data for children's heights, partitioned into classes whose parents all had the same height  $\pm 1$  inch.

## 16.2 Regression to the mean and the regression fallacy

We have all heard stories of a child coming home proudly from school with a score of 99 out of 100 on a test, and the strict parent who points out that he or she had 100 out of 100 on the last test, and “what happened to the other point?” Of course, we instinctively recognise the parent’s response as absurd. Nothing *happened* to the other point (in the sense of the child having fallen down in his or her studies); that’s just how test scores work. Sometimes they’re better, sometimes worse. It is unfair to hold someone to the standard of the last perfect score, since the next score is unlikely to be exactly the same, and there’s nowhere to go but down.

Of course, this is true of any measurements that are imperfectly correlated: If  $|r|$  is substantially less than 1, the regression equation tells us that those individuals who have extreme values of  $x$  tend to have values of  $y$  that are somewhat less extreme. If  $r = 0.5$ , those with  $x$  values that are 2 SDs above the mean will tend to have  $y$  values that are still above average, but only 1 SD above the mean. There is nothing strange about this: If we pick out those individuals who have exceptionally voluminous brains, for instance, it is not surprising that the surface areas of their brains are less extreme. While athletic ability certainly carries over from one sport to another, we do not expect the world’s finest footballers to also win gold medals in swimming. Nor does it seem odd that great composers are rarely great pianists, and vice versa.

And yet, when the  $x$  and  $y$  are successive events in time — for instance, the same child’s performance on two successive tests — there is a strong tendency to attribute causality to this imperfect correlation. Since there is a random component to performance on the test, we expect that the successive scores will be correlated, but not exactly the same. The plot of score number  $n$  against score number  $n + 1$  might look like the upper left scatterplot in Figure 15.3. If she had an average score last time, she’s likely to score about the same this time. But if she did particularly well last time, this time is likely to be less good. But consider how easy it would be to look at these results and say, “Look, she did well, and as a result she slacked off, and did worse the next time;” or “It’s good that we punish her when she does poorly on a test by not letting her go outside for a week, because that always helps her focus, and she does better the next time.” Galton noticed that children of exceptionally tall parents were closer to average than the parents were, and called this “regression to mediocrity” [Gal86].

Some other examples:

- Speed cameras tend to be sited at intersections where there have been high numbers of accidents. Some of those high numbers are certainly inherent to the site, but sometimes the site was just “unlucky” in one year. You would expect the numbers to go down the next year regardless of whether you made any changes, just as you would expect an intersection with very few accidents to have more the next year. Some experts have pointed out that this can lead to overestimating the effect of the cameras. As described in The Times 15 December 2005 [<http://www.timesonline.co.uk/tol/news/uk/article766659.ece>],

The Department for Transport [...] published a study which found that cameras saved only half as many lives as claimed. This undermines the Government's main justification for increasing speed camera penalties five-fold from 340,000 in 1997 to 1.8 million in 2003.

Safe Speed, the anti-camera campaign, has argued for years that the policy of siting cameras where there have been recent clusters of crashes makes it impossible to attribute any fall in collisions to the presence of a camera. Collisions would be expected to fall anyway as they reverted from the temporary peak to the normal rate.

The department commissioned the Department of Engineering at Liverpool University to study this effect, which is known as regression to the mean. The study concluded that most of the fall in crashes could be attributed to regression to the mean. The presence of the camera was responsible for as little as a fifth of the reduction in casualties.

The report goes on to say that “The department put the results of the study close to the bottom of a list of appendices at the back of a 160-page report which claims that cameras play an important role in saving lives.”

- Suppose you are testing a new blood pressure medication. As we have described in section 11.5, it is useful to compare the same individual's blood pressure before and after taking the medication. So we take a group of subjects, measure their blood pressure ( $x_i$ ), give them the medication for two weeks, then measure again ( $y_i$ ), and test

to see whether  $y_i - x_i$  is negative, on average. We can't give blood-pressure-lowering medication to people who have normal blood pressure, though, so we start by restricting the study to those whose first measurement  $x$  is in the hypertension range,  $x_i > 140\text{mmHg}$ . Since they are above average, and since there is significant random fluctuation in blood pressure measurements, those individuals would be expected to have lower blood pressure measurements the second time, purely by chance. If you are not careful, you will find yourself overestimating the effectiveness of the medication.

- The behavioural economists Amos Tversky and Daniel Kahneman tell the story of having taught psychology to air force flight instructors. He tried to explain to them that there was a great deal of evidence that positive reinforcement — praise for good performance — is more effective than negative reinforcement (criticism of poor performance).

After some experience with this training approach, the instructors claimed that contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try[...] Regression is inevitable in flight maneuvers because performance is not perfectly reliable and progress between successive maneuvers is slow. Hence, pilots who did exceptionally well on one trial are likely to deteriorate on the next, regardless of the instructors' reaction to the initial success. The experienced flight instructors actually discovered the regression but attributed it to the detrimental effect of positive reinforcement. **This true story illustrates a saddening aspect of the human condition.** We normally reinforce others when their behavior is good and punish them when their behavior is bad. By regression alone, therefore, they are most likely to improve after being punished and most likely to deteriorate after being rewarded. Consequently, **we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding.** [Tve82]

## 16.3 When the data don't fit the model

As part of a cross-species study of sleep behaviour [AC76] presented a table of brain and body weights for 62 different species of land mammal. We show these data in Table 16.1. We have already shown a scatterplot of brain weights against body weights in Figure 15.2(e). It is clear from looking at the data that there is some connection between brain and body weights, but it is also clear that we have some difficulty in applying the ideas of this lecture. These are based, after all, on a model in which the variables are normally distributed, so that they are distributed about in some kind of approximately oval scatterplot. The correlation is supposed to represent a summary of the relation between all x values and the corresponding y's. Here, though, the high correlation (0.93) is determined almost entirely by the elephants, which have brain and body sizes far above the mean.

### 16.3.1 Transforming the data

One approach to understanding the correlation in this data set is illustrated in Figure 15.2(f), where we have plotted log of brain weight against log of body weight. The result now looks quite a bit like our standard regression scatter plots. We can compute that the correlation of these two measures is actually even a bit higher: 0.96. Thus, we can say that 92% ( $= 0.96^2$ ) of the variance in log brain weight is explained by body weight. What is more, the regression line now seems to make some sense.

### 16.3.2 Spearman's Rank Correlation Coefficient

Taking logarithms may seem somewhat arbitrary. After all, there are a lot of ways we might have chosen to transform the data. Another approach to dealing with such blatantly nonnormal data, is to follow the same approach that we have taken in all of our nonparametric methods: We replace the raw numbers by ranks. Important: The ranking takes place within a variable. We have shown in Table 16.1, in columns 3 and 5, what the ranks are: The highest body weight — african elephant — gets 62, the next gets 61, down to the lesser short-tailed shrew, that gets rank 1. Then we start over again with the brain weights. (When two or more individuals are tied, we average the ranks.) The correlation that we compute between the ranks is called **Spearman's rank correlation coefficient**, denoted  $r_s$ . It tells us quantitatively whether high values of one variable tend to go with high values of the other, without relying on assumptions of normality or otherwise being

dominated by a few very extreme values. Since it is really just a correlation, we test it (for being different from 0) the same way we test any correlation, by comparing  $r_s \sqrt{n-2} / \sqrt{1-r_s^2}$  with the t distribution with  $n-2$  degrees of freedom.

### 16.3.3 Computing Spearman's rank correlation coefficient

There is a slightly quicker way of computing the Spearman coefficient. We first list the rankings of the two variables in two parallel rows, and let  $d_i$  be the difference between the  $x_i$  ranking and the  $y_i$  ranking. We show this calculation in Table 16.2. We then have  $D = \sum d_i^2 = 1846.5$ . We have then the formula

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 0.954.$$

This is exactly what we get when we compute the correlation between the two lists of ranks, as described in section 15.4.

Species	Body weight (kg)	Body rank	Brain weight (g)	Brain rank
African elephant	6654.00	62	5712.00	62
African giant pouched rat	1.00	21	6.60	22
Arctic Fox	3.38	32	44.50	37
Arctic ground squirrel	0.92	20	5.70	19
Asian elephant	2547.00	61	4603.00	61
Baboon	10.55	42	179.50	50
Big brown bat	0.02	4	0.30	3
Brazilian tapir	160.00	53	169.00	47
Cat	3.30	31	25.60	35
Chimpanzee	52.16	47	440.00	56
Chinchilla	0.42	14	6.40	21
Cow	465.00	58	423.00	55
Desert hedgehog	0.55	16	2.40	10
Donkey	187.10	54	419.00	54
Eastern American mole	0.07	7	1.20	8
Echidna	3.00	30	25.00	34
European hedgehog	0.79	18	3.50	14
Galago	0.20	12	5.00	17
Genet	1.41	25	17.50	32
Giant armadillo	60.00	49	81.00	41
Giraffe	529.00	60	680.00	59
Goat	27.66	44	115.00	44
Golden hamster	0.12	10	1.00	6
Gorilla	207.00	56	406.00	53
Gray seal	85.00	51	325.00	52
Gray wolf	36.33	46	119.50	45
Ground squirrel	0.10	8	4.00	16
Guinea pig	1.04	22	5.50	18
Horse	521.00	59	655.00	58
Jaguar	100.00	52	157.00	46
Kangaroo	35.00	45	56.00	39
Lesser short-tailed shrew	0.01	1	0.14	1
Little brown bat	0.01	2	0.25	2
Man	62.00	50	1320.00	60
Mole rat	0.12	11	3.00	13
Mountain beaver	1.35	23	8.10	23
Mouse	0.02	4	0.40	5
Musk shrew	0.05	5	0.33	4
N. American opossum	1.70	27	6.30	20
Nine-banded armadillo	3.50	34	10.80	24
Okapi	250.00	57	490.00	57
Owl monkey	0.48	15	15.50	30
Patas monkey	10.00	41	115.00	44
Phanlanger	1.62	26	11.40	25
Pig	192.00	55	180.00	51
Rabbit	2.50	29	12.10	26
Raccoon	4.29	39	39.20	36
Rat	0.28	13	1.90	9
Red fox	4.24	38	50.40	38
Rhesus monkey	6.80	40	179.00	49
Rock hyrax (Hetero. b)	0.75	17	12.30	28
Rock hyrax (Procavia hab)	3.60	35	21.00	33
Roe deer	14.83	43	98.20	42
Sheep	55.50	48	175.00	48
Slow loris	1.40	24	12.50	29
Star nosed mole	0.06	6	1.00	6
Tenrec	0.90	19	2.60	12
Tree hyrax	2.00	28	12.30	28
Tree shrew	0.10	9	2.50	11
Vervet	4.19	37	58.00	40
Water opossum	3.50	34	3.90	15
Yellow-bellied marmot	4.05	36	17.00	31
mean	199		243	
SD	899	930		

Table 16.1: Brain and body weights for 62 different land mammal species. Available at <http://lib.stat.cmu.edu/datasets/sleep>, and as the object `mammals` in the statistical language `sR`.

body	62	21	32	20	61	42	4	53	31	47	14	58	16	54	7	30	18	12	25	49	60
brain	62	22	37	19	61	50	3	47	35	56	21	55	10	54	8	34	14	17	32	41	59
diff.	0	1	5	1	0	8	0	6	4	9	7	3	6	0	1	4	4	5	7	8	1
body	44	10	56	51	46	8	22	59	52	45	1	2	50	11	23	4	5	27	34	57	15
brain	44	6	53	52	45	16	18	58	46	39	1	2	60	13	23	5	4	20	24	57	30
diff.	0	4	3	1	1	8	4	1	6	6	0	0	10	2	0	2	1	7	10	0	15
body	41	26	55	29	39	13	38	40	17	35	43	48	24	6	19	28	9	37	34	36	
brain	44	25	51	26	36	9	38	49	28	33	42	48	29	6	12	28	11	40	15	31	
diff.	2	1	4	3	3	4	0	9	10	2	1	0	5	0	7	0	2	3	18	5	

Table 16.2: Ranks for body and brain weights for 62 mammal species, from Table 16.1, and the difference in ranks between body and brain weights.

# Bibliography

- [Abr78] Sidney Abraham. *Total serum cholesterol levels of children, 4-17 years, United States, 1971-1974*. Number 207 in Vital and Health statistics: Series 11, Data from the National Health Survey. National Center for Health Statistics, 1978.
- [AC76] Truett Allison and Domenic V. Cicchetti. Sleep in mammals: Ecological and constitutional correlates. *Science*, 194:732–4, November 12 1976.
- [BH84] Joel G. Breman and Norman S. Hayner. Guillain-Barré Syndrome and its relationship to swine influenza vaccination in Michigan, 1976-1977. *American Journal of Epidemiology*, 119(6):880–9, 1984.
- [Bia05] Carl Bialik. When it comes to donations, polls don't tell the whole story. *The Wall Street Journal*, 2005.
- [BS] Norman R. Brown and Robert C. Sinclair. Estimating number of lifetime sexual partners: Men and women do it differently. <http://www.ualberta.ca/~nrbrown/pubs/BrownSinclair1999.pdf>.
- [Edw58] A. W. F. Edwards. An analysis of Geissler's data on the human sex ratio. *Annals of Human Genetics*, 23(1):6–15, 1958.
- [Fel71] William Feller. *An Introduction to Probability and its Applications*, volume 2. John Wiley & Sons, New York, 1971.
- [FG00] Boris Freidlin and Joseph L. Gastwirth. Should the median test be retired from general use? *The American Statistician*, 54(3):161–4, August 2000.

- [FPP98] David Freedman, Robert Pisani, and Roger Purves. *Statistics*. Norton, 3 edition, 1998.
- [Gal86] Francis Galton. Regression toward mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15:246–63, 1886.
- [Hal90] Anders Hald. *History of Probability and Statistics and their applications before 1750*. John Wiley & Sons, New York, 1990.
- [HDL<sup>+</sup>94] D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski. *A Handbook of small data sets*. Chapman & Hall, 1994.
- [HREG97] Anthony Hill, Julian Roberts, Paul Ewings, and David Gunnell. Non-response bias in a lifestyle survey. *Journal of Public Health*, 19(2), 1997.
- [LA98] J. K. Lindsey and P. M. E. Altham. Analysis of the human sex ratio by using overdispersion models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(1):149–157, 1998.
- [Lev73] P. H. Levine. An acute effect of cigarette smoking on platelet function. *Circulation*, 48:619–23, 1973.
- [Lov71] C. Owen Lovejoy. Methods for the detection of census error in palaeodemography. *American Anthropologist, New Series*, 73(1):101–9, February 1971.
- [LSS73] Judson R. Landis, Daryl Sullivan, and Joseph Sheley. Feminist attitudes as related to sex of the interviewer. *The Pacific Sociological Review*, 16(3):305–14, July 1973.
- [MDF<sup>+</sup>81] G. S. May, D. L. DeMets, L. M. Friedman, C. Furberg, and E. Passamani. The randomized clinical trial: bias in analysis. *Circulation*, 64:669–673, 1981.
- [MM98] David S. Moore and George P. McCabe. *Introduction to the Practice of Statistics*. W. H. Freeman, New York, 3rd edition, 1998.
- [MMSR02] Erik Lykke Mortensen, Kim Fleischer Michaelsen, Stephanie A. Sanders, and June Machover Reinisch. The association between duration of breastfeeding and adult intelligence. *JAMA*, 287(18):2365–71, May 8 2002.

- 
- [Mou98] Richard Francis Mould. *Introductory medical statistics*. CRC Press, 3 edition, 1998.
  - [MS84] Marc Mangel and Francisco J. Samaniego. Abraham wald's work on aircraft survivability. *Journal of the American Statistical Association*, 79(386):259–267, 1984.
  - [NNGT02] Mark J. Nieuwenhuijsen, Kate Northstone, Jean Golding, and The ALSPAC Study Team. Swimming and birth weight. *Epidemiology*, 13(6):725–8, 2002.
  - [Ric95] John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.
  - [RS90] Luigi M. Ricciardi and Shunsuke Sato. Diffusion processes and first-passage-time problems. In Luigi M. Ricciardi, editor, *Lectures in applied mathematics and informatics*, pages 206–285. Manchester Univ. Press, Manchester, 1990.
  - [RS02] Fred L. Ramsey and Daniel W. Schafer. *The Statistical Sleuth: A course in methods of data analysis*. Duxbury Press, 2nd edition, 2002.
  - [RSKG85] S. Rosenbaum, R. K. Skinner, I. B. Knight, and J. S. Garrow. A survey of heights and weights of adults in Great Britain, 1980. *Annals of Human Biology*, 12(2):115–27, 1985.
  - [SCB06] Emad Salib and Mario Cortina-Borja. Effect of month of birth on the risk of suicide. *British Journal of Psychiatry*, 188:416–22, 2006.
  - [SCT<sup>+</sup>90] R. L. Suddath, G. W. Christison, E. F. Torrey, M. F. Casanova, and D. R. Weinberger. Anatomical abnormalities in the brains of monozygotic twins discordant for schizophrenia. *New England Journal of Medicine*, 322(12):789–94, March 22 1990.
  - [SKK91] L. Sheppard, A. R. Kristal, and L. H. Kushi. Weight loss in women participating in a randomized trial of low-fat diets. *The American Journal of Clinical Nutrition*, 54:821–8, 1991.
  - [TK81] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–8, January 30 1981.

- [TLG<sup>+</sup>98] M. J. Tramo, W. C. Loftus, R. L. Green, T. A. Stukel, J. B. Weaver, and M. S. Gazzaniga. Brain size, head size, and intelligence quotient in monozygotic twins. *Neurology*, 50:1246–52, 1998.
- [TPCE06] Anna Thorson, Max Petzold, Nguyen Thi Kim Chuc, and Karl Ekdahl. Is exposure to sick or dead poultry associated with flulike illness?: A population-based study from a rural area in Vietnam with outbreaks of highly pathogenic avian influenza. *Archives of Internal Medicine*, 166:119–23, 2006.
- [Tve82] Amos Tversky. On the psychology of prediction. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgment under uncertainty: heuristics and biases*. Cambridge University Press, 1982.
- [vBFHL04] Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley. *Biostatistics: A methodology for the health sciences*. Wiley-IEEE, 2nd edition, 2004.
- [ZZK72] Philip R. Zelazo, Nancy Ann Zelazo, and Sarah Kolb. “walking” in the newborn. *Science*, 176(4032):314–5, April 21 1972.