



Biostatistics for Biomedical Research
Vanderbilt Institute for Clinical and Translational Research
Edge for Scholars
Department of Biostatistics

Biostatistics for Biomedical Research

Frank E Harrell Jr
James C Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
f.harrell@vanderbilt.edu
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/ClinStat
Questions/discussions/topic suggestions:
datamethods.org/t/bbr-video-course
Web course: hbiostat.org/bbr
R code in text: hbiostat.org/bbr/code.zip
Glossary of statistical terms: bit.ly/stat-glossary
Blog: fharrell.com
@f2harrell #bbrcourse

Contents

1 R	1-1
1.1 Background	1-1
1.2 Learning R	1-3
1.3 Setting up R	1-5
1.4 Using R Markdown	1-7
1.5 Debugging R Code	1-9
1.6 Importing Other Datasets	1-10
1.7 Suggestions for Initial Data Look	1-15
1.8 Operating on Data Frames	1-16
2 Algebra Review	2-1
2.1 Overview	2-1
2.2 Some Resources	2-4
3 General Overview of Biostatistics	3-1
3.1 What is Biostatistics?	3-2
3.2 What Can Statistics Do?	3-4
3.3 Types of Data Analysis and Inference	3-8
3.4 Types of Measurements by Their Role in the Study	3-10
3.5 Types of Measurements According to Coding	3-12
3.6 Choose Y to Maximize Statistical Information, Power, and Interpretability	3-13

3.7 Preprocessing	3-16
3.8 Random Variables	3-17
3.9 Probability	3-18
4 Descriptive Statistics, Distributions, and Graphics	4-1
4.1 Distributions	4-1
4.2 Descriptive Statistics	4-7
4.3 Graphics	4-10
4.4 Tables	4-35
5 Statistical Inference	5-1
5.1 Overview	5-1
5.2 Hypotheses	5-7
5.3 Branches of Statistics	5-9
5.4 Errors in Hypothesis Testing: <i>P</i> Values	5-14
5.5 Interval Estimation	5-19
5.6 One Sample Test for Mean	5-22
5.7 One Sample Method for a Probability	5-41
5.8 Paired Data and One-Sample Tests	5-48
5.9 Two Sample Test for Means	5-49
5.10 Comprehensive Example: Two sample <i>t</i> -test	5-57
5.11 The Problem with Hypothesis Tests and <i>P</i> -values Revisited	5-68
5.12 Study Design Considerations	5-71
5.13 One-Sample <i>t</i> -Test Revisited	5-78
5.14 Comprehensive Example: Crossover Design and Analysis	5-81
6 Comparing Two Proportions	6-1
6.1 Overview	6-1

6.2	Normal-Approximation Test	6-2
6.3	χ^2 Test	6-4
6.4	Fisher's Exact Test	6-6
6.5	Sample Size and Power for Comparing Two Independent Samples	6-7
6.6	Confidence Interval	6-9
6.7	Sample Size for a Given Precision	6-10
6.8	Relative Effect Measures	6-12
6.9	Comprehensive example	6-14
6.10	Logistic Regression for Comparing Proportions	6-18
7	Nonparametric Statistical Tests	7-1
7.1	When to use non-parametric methods	7-1
7.2	One Sample Test: Wilcoxon Signed-Rank	7-4
7.3	Two Sample Test: Wilcoxon–Mann–Whitney	7-9
7.4	Confidence Intervals for Medians and Their Differences	7-16
7.5	Strategy	7-19
7.6	Generalization of the Wilcoxon/Kruskal-Wallis Test	7-20
7.7	Checking Assumptions of the Wilcoxon Test	7-28
7.8	Power and Sample Size	7-30
7.9	Bayesian Proportional Odds Model	7-39
8	Correlation	8-1
8.1	Overview	8-1
8.2	Pearson's correlation coefficient	8-2
8.3	Spearman's Rank Correlation	8-4
8.4	Correlation Examples	8-5
8.5	Correlation and Agreement	8-9
8.6	Avoiding Overinterpretation	8-16

9 Introduction to the R rms Package: The Linear Model	9-1
9.1 Formula Language and Fitting Function	9-2
9.2 Operating on the Fit Object	9-3
9.3 The rms datadist Function	9-4
9.4 Short Example	9-5
9.5 Operating on Residuals	9-9
9.6 Plotting Partial Effects	9-10
9.7 Nomograms: Overall Depiction of Fitted Models	9-16
9.8 Getting Predicted Values	9-19
9.9 ANOVA	9-21
10 Simple and Multiple Regression Models	10-1
10.1 Stratification vs. Matching vs. Regression	10-4
10.2 Purposes of Statistical Models	10-6
10.3 Advantages of Modeling	10-7
10.4 Nonparametric Regression	10-8
10.5 Simple Linear Regression	10-10
10.6 Proper Transformations and Percentiling	10-24
10.7 Multiple Linear Regression	10-29
10.8 Multiple Regression with a Binary Predictor	10-40
10.9 The Correlation Coefficient Revisited	10-42
10.10 Using Regression for ANOVA	10-44
10.11 Internal vs. External Model Validation	10-54
11 Multiple Groups	11-1
11.1 Examples	11-1
11.2 The k -Sample Problem	11-2
11.3 Parametric ANOVA	11-3

11.4 Why All These Distributions?	11-9
11.5 Software and Data Layout	11-11
11.6 Comparing Specific Groups	11-12
11.7 Non-Parametric ANOVA: Kruskal-Wallis Test	11-13
11.8 Two-Way ANOVA	11-15
11.9 Analysis of Covariance	11-17
11.10 Multiple Comparisons	11-18
12 Statistical Inference Review	12-1
12.1 Types of Analyses	12-2
12.2 Covariate-Unadjusted Analyses	12-3
12.3 Covariate-Adjusted Analyses	12-5
13 Analysis of Covariance in Randomized Studies	13-1
13.1 Covariate Adjustment in Linear Models	13-2
13.2 Covariate Adjustment in Nonlinear Models	13-3
13.3 Cox / Log-Rank Test for Time to Event	13-7
13.4 Why are Adjusted Estimates Right?	13-9
13.5 How Many Covariates to Use?	13-10
13.6 Differential and Absolute Treatment Effects	13-11
13.7 Cost-Effectiveness Ratios	13-31
13.8 Treatment Contrasts for Multi-Site Randomized Trials	13-33
13.9 Statistical Plan for Randomized Trials	13-34
13.10 Summary	13-37
13.11 Notes	13-38
14 Transformations, Measuring Change, and Regression to the Mean	14-1
14.1 Transformations	14-1

14.2 Logarithmic Transformation	14-2
14.3 Analysis of Paired Observations	14-6
14.4 What's Wrong with Change in General?	14-7
14.5 What's Wrong with Percent Change?	14-13
14.6 Objective Method for Choosing Effect Measure	14-15
14.7 Example Analysis: Paired Observations	14-16
14.8 Regression to the Mean	14-19
15 Serial Data	15-1
15.1 Introduction	15-1
15.2 Analysis Options	15-3
15.3 Case Study	15-6
16 Analysis of Observer Variability and Measurement Agreement	16-1
16.1 Intra- and Inter-observer Disagreement	16-1
16.2 Comparison of Measurements with a Standard	16-9
16.3 Special Case: Assessing Agreement In Two Binary Variables	16-10
16.4 Problems	16-13
16.5 References	16-14
17 Modeling for Observational Treatment Comparisons	17-1
17.1 Propensity Score	17-1
17.2 Assessing Treatment Effect	17-3
17.3 Recommended Statistical Analysis Plan	17-5
17.4 Reasons for Failure of Propensity Analysis	17-7
17.5 Sensitivity Analysis	17-8
17.6 Reasons To Not Use Propensity Analysis	17-9
17.7 Further Reading	17-10

18 Information Loss	18-1
18.1 Information & Decision Making	18-2
18.2 Ignoring Readily Measured Variables	18-4
18.3 Categorization: Partial Use of Information	18-6
18.4 Problems with Classification of Predictions	18-18
18.5 Value of Continuous Markers	18-20
18.6 Harm from Ignoring Information	18-25
18.7 Case Study in Faulty Dichotomization of a Clinical Outcome: Statistical and Ethical Concerns in Clinical Trials for Crohn's Disease	18-30
18.8 Information May Sometimes Be Costly	18-36
19 Diagnosis	19-1
19.1 Problems with Traditional Indexes of Diagnostic Utility	19-3
19.2 Problems with ROC Curves and Cutoffs	19-7
19.3 Optimum Individual Decision Making and Forward Risk	19-8
19.4 Diagnostic Risk Modeling	19-10
19.5 Assessing Diagnostic Yield	19-14
19.6 Assessing Absolute Diagnostic Yield: Cohort Study	19-15
19.7 Assessing Diagnostic Yield: Case-Control & Other Oversampling Designs	19-17
19.8 Example: Diagnosis of Coronary Artery Disease (CAD): Test = Total Cholesterol	19-18
19.9 Summary	19-20
20 Challenges of Analyzing High-Dimensional Data	20-1
20.1 Background	20-2
20.2 Global Analytic Approaches	20-5
20.3 Simulated Examples	20-10
21 Reproducible Research	21-1
21.1 Non-reproducible Research	21-2

21.2 General Importance of Sound Methodology	21-3
21.3 System Forces	21-7
21.4 Strong Inference	21-8
21.5 Pre-Specified Analytic Plans	21-9
21.6 Summary	21-10
21.7 Software	21-11
21.8 Further Reading	21-16
Bibliography	22-1

Changes Since 2017-01-01

Section	Date Modified	Description
	2017-01-25	Began adding slack discussion keys
20.3	2017-02-01	Suggested stat plan wording for ranking genes
13	2017-03-09	Slack discussion keys added, new intro
6.9.4	2017-05-31	Upper CL for odds ratio was wrong
13.6.1	2017-06-22	New examples of displaying HTE
5.6.3	2017-09-12	Added MMOE for estimating odds
21.1	2017-09-20	Added new comprehensive ref on repro res
14	2017-11-12	Several improvements in analysis of change
7.4	2017-12-01	Clarified difference in medians vs. H-L
3.5	2017-12-24	New section on statistical information by variable type
19.3	2017-12-29	New web site wonderfully illustrating optimum Bayes decisions
8.5.1	2017-12-29	New text and reference: controversy over Bland-Altman
13.1	2017-12-31	Added Bland-Altman reference on impropriety of testing for baseline balance in RCT
21	2018-01-08	Added material from personalized med talk
	2018-01-28	code.zip updated, moved to fharrell.com/code
18.3.1	2018-04-10	Added example of categorizing height and weight when assessing importance of BMI
10.11.1	2018-04-13	New subsection on choice of validation methods
4.2.2	2018-04-14	Added Gini's mean difference
3.8	2018-04-14	New section - probability overview
5.8.2	2019-06-21	Example of needed observed differences in means
1.2	2019-06-29	Link to Norm Matloff material for learning R
13	2019-08-11	Added link to Senn article, fixed plotting contrasts as.data.frame
3	2019-08-25	many
4.3	2019-08-30	reference RCTGraphics wiki
3.2.1	2019-09-16	new section
3.8	2019-10-10	added McElreath's terminology
3.9	2019-10-10	added baseball batting average example
3.2.1.1	2019-10-15	new section on planning observational research
4.3	2019-10-15	new links, more info about interactive graphics
4.5	2019-10-16	moved to 4.3.4
4.3.4	2019-10-16	new section on semi-interactive graphics
4.4	2019-10-19	new opinions about tables
5	2019-10-22	many changes for BBR course
5.1.1	2019-10-24	new section on central limit theorem
5.5	2019-10-27	new section on interval estimation
5.6	2019-11-09	added Bayes
5.9	2019-12-02	added Bayes, many other changes
	2019-12-07	added hyperlinks for YouTube BBR sessions and datamethods.org
5.12	2019-12-09	many
6	2019-12-11	added frequentist and Bayesian logistic model
7	2019-12-30	several minor changes
7.6-7.8	2020-01-19	new sections on PO model and power
8	2020-01-25	new sections on bootstrapping correlation matrices
14.4.1	2020-02-03	Added S Senn change from baseline graph

Blue symbols in the right margin starting with ABD designate section numbers (and occasionally page numbers preceded by *p*) in *The Analysis of Biological Data, Second Edition* by MC Whitlock and D Schluter, Greenwood Village CO, Roberts and Company, 2015. Likewise, right blue symbols starting with RMS designate section numbers in *Regression Modeling Strategies, 2nd ed.* by FE Harrell, Springer, 2015.



in the right margin indicates a hyperlink to a YouTube video related to the subject.

There is a YouTube channel (BBRcourse) for these notes at bit.ly/yt-bbr or by searching for YouTube channel BBRcourse. A discussion board about the overall course is at datamethods.org/t/bbr-video-course. You can go directly to a YouTube video for BBR Session n by going to bit.ly/yt-bbrn.

 in the right margin is a hyperlink to the discussion topic in datamethods.org devoted to the specific YouTube video session. You can go directly to the discussion about session n by going to bit.ly/datamethods-bbrn. Some of the session on YouTube also had live chat which you can select to replay while watching the video.

 in the right margin is a hyperlink to an audio file^a elaborating on the notes. Red letters and numbers in the right margin are cues referred to within the audio recordings.

[Here](#) is a link to the playlist for all audio files in these notes.

Rotated boxed blue text in the right margin at the start of a section represents the mnemonic key for linking to discussions about that section in vbiostatcourse.slack.com channel #bbr. Anyone starting a new discussion about a topic related to the section should include the mnemonic somewhere in the posting, and the posting should be marked to slack as threaded. The mnemonic in the right margin is also a hyperlink to a search in the bbr channel for messages containing the mnemonic. When you click on it the relevant messages will appear in the search results on the right side of the slack browser window. Note: Slack is not being used for new discussions. Use datamethods.org topics devoted to BBR video sessions instead, or open a new topic for topics that are not part of the BBR web course.

howto

Members of the slack group can also create submnemonics for subsections or other narrower-scope parts of the notes. When creating keys “on the fly,” use names of the form chapterkey-sectionkey-yourkey where sectionkey is defined in the notes. That way a search on chapterkey-sectionkey will also bring up notes related to yourkey.

Several longer and more discussed subsections in the text have already been given short keys in these notes.

Beginning in 2019, the use of slack for questions, answers, and discussion is no longer recommended. Instead use datamethods.org where separate special topics are created for some of the chapters in these notes.

[blog](#) in the right margin is a link to a blog entry that further discusses the topic.

^aThe first time you click on one of these, some browsers download the audio file and give you the opportunity to right click to open the file on your local audio player, then the browser asks if you always want to open files of this type. Select “yes”.

Chapter 1

R

1.1

Background

Computer code shown throughout these notes is R¹⁰³. R is free and is the most widely used statistical software in the world. It has the best graphics, statistical modeling, nonparametric methods, survival analysis, clinical trials methods, and data manipulation capabilities. R has the most comprehensive genomics analysis packages and has advanced capabilities for reproducible analysis and reporting. R also has an excellent graphical front-end RStudio (rstudio.org) that has the identical look and feel on all operating systems and via a web browser. Part of R's appeal is the thousands of add-on packages available (at <http://cran.r-project.org/web/packages>), which exist because it is easy to add to R. Many of the add-on packages are specialty packages for biomedical research including packages for such widely diverse areas as

- interfacing R to REDCap (2 packages)
- interactive design of adaptive clinical trials
- analyzing accelerometer data
- flow cytometry
- genomics
- analyzing ICD9 codes and computing comorbidity indexes

- downloading all annotated NHANES datasets
- interfacing to [clinicaltrials.gov](#)
- detecting whether a dataset contains personal identifiers of human subjects
- analysis of early phase cardiovascular drug safety studies

The main R web site is [www.r-project.org](#).

1.2

Learning R

Start with *R Tutorials* at <http://www.r-bloggers.com/how-to-learn-r-2>, *R Programming Tutorials* from Mike Marin at <https://www.youtube.com/user/marinstatlecture> or *Fast Lane to Learning R* at <https://github.com//matloff/fasteR> by Norm Matloff. Or look at swirlstats.com for an interactive way to learn R. Those who have used SPSS or SAS before will profit from *R for SAS and SPSS Users* by Robert Muenchen. A current list of R books on amazon.com may be found at <http://amzn.to/15URiF6>. <https://stats.idre.ucla.edu/r/> and http://www.introductoryr.co.uk/R_Resources_for_Beginners.html are useful web sites. An excellent resource is *R for Data Science* by Grolemund and Wickham. See also *R in Action, second ed.* by Robert Kabacoff. The online open source book on statistical modeling by Legler and Roback at <https://bookdown.org/roback/bookdown-bysh> contains a lot of R code. Jenny Bryan's *STAT 545 Data Wrangling, Exploration, and Analysis with R* course at stat545.com is an excellent resource. <http://stackoverflow.com/tags/r> is the best place for asking questions about the language and for learning from answers to past questions asked (see also the R-help email list).

Three of the best ways to learn how to analyze data in R quickly are

1. Avoid importing and manipulating data, instead using the R `load` function to load datasets that are already annotated and analysis-ready (see Section 1.6 for information about importing your own datasets)
2. Use example R scripts as analysis templates
3. Use RStudio (rstudio.org) to run R

On the first approach, the R `Hmisc` package's `getHdata` function finds datasets on the Vanderbilt Biostatistics DataSets wiki, downloads them, and `load()`s them in your R session. These notes use only datasets available via this mechanism. These datasets are fully annotated with variable labels and units of measurements for many of the continuous variables. Concerning analysis scripts, Vanderbilt Biostatistics has collected template analysis scripts on <https://github.com/harrelfe/rscripts>^a and the R `Hmisc` package has a function `getRs` to download these scripts and to automatically

^agithub has outstanding version control and issue reporting/tracking. It greatly facilitates the contribution of new scripts by users, which are most welcomed. Contact f.harrell@vanderbilt.edu if you have scripts to contribute or suggestions for existing scripts.

populate an RStudio script editor window with the script. Many of the scripts are in RMarkdown format for use with the R `knitr` package to allow mixing of text and R code to make reproducible reports. `knitr` is described in Section [21.7](#).

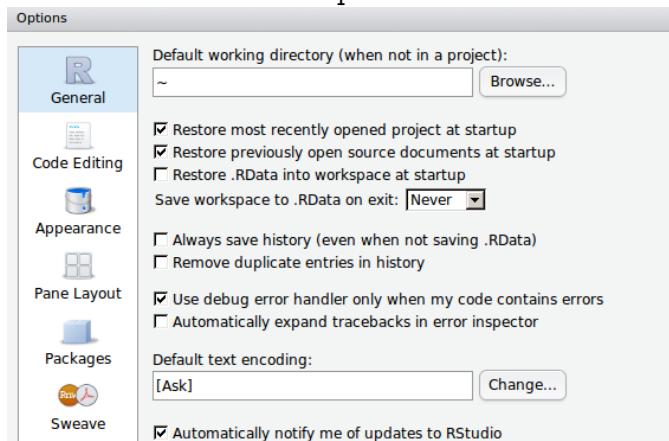
The RMarkdown scripts accessed through `getRs` use a template that makes the result part of a reproducible research process by documenting the versions of R and attached packages at the end of the report. Some of the scripts make use of the `knitrSet` function in the `Hmisc` package. When running Rmarkdown, call `knitrSet(lang='markdown')`. `knitrSet` gets rid of `##` at the start of R output lines, and makes it easy to specify things like figure sizes in knitr chunk headers. It also causes annoying messages such as those generated from attaching R packages to be put in a separate file `messages.txt` rather than in the report.

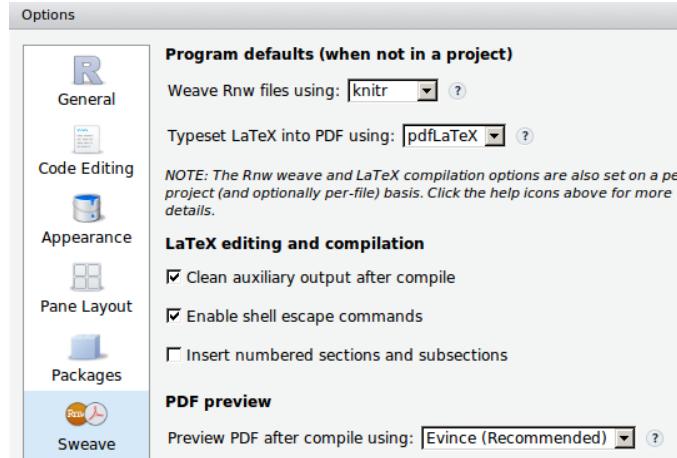
1.3

Setting up R

Before running examples in these notes and R markdown example scripts, you need to do the following:

1. Make sure your operating system is up to date enough to run the most current version of R at www.r-project.org. For Mac you must have OS X Maverick or later.
2. Install R from www.r-project.org or upgrade your installation of R to the latest version.
3. Install RStudio from rstudio.org or update your RStudio to the latest version.
4. Run RStudio and get it to install the packages that allow Rmarkdown to run, by clicking on File ... New File ... R Markdown. Make sure that the knitr package is installed.
5. Have RStudio install the Hmisc and rms packages (which will make RStudio install several other packages). For packages you had installed previously, make sure you update them to have the latest versions.
6. Configure RStudio Tools ... Global Options to match the images below





Here are some examples of how `getRs` is used once you load the `Hmisc` package using a menu or by typing `require(Hmisc)` or `library(Hmisc)` in the console.

```
require(Hmisc)          # do this once per session (or library(Hmisc))
options(url.method='libcurl')    # sometimes needed if using Windows
getRs()                # list available scripts
getRs(browser='browser')  # open scripts contents in your web browser
scripts <- getRs()      # store directory of scripts in an object that can easily
                        # be viewed on demand in RStudio (right upper pane)
getRs('introda.r')     # download introda.r and open in script editor
getRs(cats=TRUE)        # list available major and minor categories
categories <- getRs(cats=TRUE) # store results in a list for later viewing
getRs(cats='reg')       # list all scripts in a major category containing 'reg'
getRs('importREDCap.r', put='source') # source() to define a function
```

You can also point your browser to <https://github.com/harrelfe/rscripts/blob/master/contents.md> to see the available scripts and categories, and to be able to click on links to see html report output.

To get started using R in RStudio to create reproducible annotated reports, finish the above configuration instructions and type the following in the RStudio console: `getRs('descriptives.Rmd')`. The above the script editor window click on Knit HTML.

1.4

Using R Markdown

See http://kbroman.org/knitr_knutshell/pages/Rmarkdown.html and print the R Markdown cheat sheet from <http://www.rstudio.com/resources/cheatsheets>.

To make the code listing pretty, put this chunk at the top of your report. `echo=FALSE` suppresses this setup chunk from printing in the report.

```
```{r setup,echo=FALSE}
require(Hmisc)
knitrSet('myreport', lang='markdown')
```
```

The argument `'myreport'` is replaced with a string to use as a prefix to all the graphics file names, making each report in your working directory use graphics file names that do not collide with each other. For example if your report is called `acidity_analysis.Rmd` you might specify `knitrSet('acidity_analysis.Rmd', lang='markdown')`. There are many other options to `knitrSet`. A commonly used one is `width=n` to specify a line width for printing code of `n` letters. The default is 61. You can also specify `echo`,`results`, and other options. Type `?knitrSet` for help.

The R `knitr` package is used to run the markdown report and insert graphics and text output into the report at appropriate slots. It is best to specify a name for each chunk, and you must use unique names. Each R code chunk must begin exactly with ````{r ...}` and the chunk name is the first set of characters that appear after the space after `r`. Here are some example chunk headers. Chunk names must not contain a space.

```
```{r descriptives}
```{r anova}
```{r anova-y1}
```{r anova_y1}
```{r acidity_plot}
```{r plot_residuals,top=1}
```{r plot_residuals,mfrow=c(2,2),left=1,top=1,rt=1,bot=1}
```{r plot-residuals,w=5,h=4}
```

Chunk options that were used above are:

Options	Description
top=1	Leave an extra line of space at top of graph for title
mfrow=c(2,2)	Use base graphics and put the next 4 plots into a single figure with 2 rows, 2 columns
left=1,rt=1,bot=1	Leave one extra line for margin for left, right, bottom of figure
w=5,h=4	Make the figure larger than the default that <code>knitrSet</code> uses (4 inch width by 3 inch height)

Always having a chunk name also allows easy navigation of chunks by clicking to the right of the green c at the bottom of your script. This will show the names of all chunks and you can click on one to go there.

1.5

Debugging R Code

When using RStudio and `knitr` as with `RMarkdown`, it is best to debug your commands a piece at a time. The fastest way to do this is to go to some line inside your first chunk and click the green `C` just above and to the right of your script. Click on `Run Current Chunk` then on `Run Next Chunk`. Shortcut keys for these are `Ctrl+Alt+C` and `Ctrl+Alt+N` (`Command+Option+C` and `Command+Option+N` for Mac). You can also click on a single line of code and run it by clicking on `Run`.

Whenever you get a strange execution error it is sometimes helpful to show the history of all the function calls leading to that error. This is done by typing `traceback()` at the command prompt.

1.6

Importing Other Datasets

Most of the work of getting some data sources ready for analysis involves reshaping datasets from wide to tall and thin, recoding variables, and merging multiple datasets. R has first-class capabilities for all of these tasks but this part of R is harder to learn, partly because there are so many ways to accomplish these tasks in R. Getting good variable names, variable labels, and value labels, can also be tedious but is highly worth the time investment.

1.6.1

Stata and SPSS

If you have Stata or SPSS files that are already shaped correctly and have variable labels and value labels, the R `Hmisc` package's `stata.get` and `spss.get` functions will produce fully annotated ready-to-analyze R data frames.

1.6.2

REDCap

REDCap exports data to R, and Biostatistics has an R function to make the import process much easier. Here is an example showing how to fetch and use the function. In this example, the user did not provide the name of the file to import but rather let the function find the last created REDCap export files in the current working directory.

```
require(Hmisc)
getRs('importREDCap.r', put='source') # source() code to define function
mydata <- importREDCap() # by default operates on last downloaded export
Save(mydata) # Hmisc function to create mydata.rda in compressed format
```

Advanced users can hook into REDCap dynamically with R to avoid the need to export/import.

1.6.3

Spreadsheets

If you have a properly formatted `csv` file (e.g., exported from a spreadsheet), the `Hmisc` `csv.get` function will read it, facilitate handling of date variables, convert column names to legal R names, and save the original column names as variable labels.

Here is an example of importing a `csv` file into R. First of all make sure your spreadsheet is a “spreadsheet from heaven” and not a “spreadsheet from hell” by reading <http://biostat.mc.vanderbilt.edu/DataTransmissionProcedures>. Then use your spreadsheet software to export a single worksheet to create a `csv` file. Small `csv` files may be pasted into your R script as is done in the following example, but in most cases you will call `csv.get` with an external file name as the first argument.

```
# What is between data <- .. and ') is exactly like an external .csv file
data <- textConnection('
Age in Years,sex,race,visit date,m/s
23,m,w,10/21/2014,1.1
14,f,b,10/22/2014,1.3
,f,w,10/15/2014,1.7
')
require(Hmisc)
d <- csv.get(data, lowernames=TRUE, datevars='visit.date',
              dateformat='%m/%d/%Y')
close(data)
# lowernames=TRUE: convert variable names to lower case
# Omit dateformat if dates are in YYYY-MM-DD format
contents(d)
```

```
Data frame:d      3 observations and 5 variables      Maximum # NAs:1
```

	Labels	Levels	Class	Storage	NAs
age.in.years	Age in Years		integer	integer	1
sex	sex	2	integer	integer	0
race	race	2	integer	integer	0
visit.date	visit date		Date	double	0
m.s	m/s		numeric	double	0

```
+-----+-----+
| Variable | Levels |
+-----+-----+
|   sex    | f,m   |
+-----+-----+
|   race   | b,w   |
+-----+-----+
```

```
d
```

```

age.in.years sex race visit.date m.s
1           23   m     w 2014-10-21 1.1
2           14   f     b 2014-10-22 1.3
3          NA   f     w 2014-10-15 1.7

```

In the contents output above you can see that the original column names have been placed in the variable labels, and the new names have periods in place of blanks or a slash, since these characters are illegal in R names.

You can have as the first argument to `csv.get` not only a file name but a URL to a file on the web. You can also specify delimiters other than commas.

Also see the excellent tutorial on importing from Excel found at <http://www.r-bloggers.com/r-tutorial-on-reading-and-importing-excel-files-into-r>.

The `Hmisc upData` function may be used to rename variables and provide variable and value labels and units of measurement. Here is another example where there is a junk variable to delete after importing, and a categorical variable is coded as integers and need to have value labels defined after importing. We show how `csv.get` automatically renamed one illegal (to R) variable name, how to redefine a variable label, and how to define the value labels. Suppose that file `test.csv` exists in our project directory and has the following contents.

```

age,sys bp,sex,junk,state
23,140,male,1,1
42,131,female,2,1
45,127,female,3,2
37,141,male,4,2

```

Now import and modify the file.

```

require(Hmisc)
d <- csv.get('test.csv')
names(d) # show names after modification by csv.get

```

```
[1] "age"      "sys.bp"    "sex"       "junk"      "state"
```

```
contents(d) # show labels created by csv.get
```

```
Data frame:d      4 observations and 5 variables      Maximum # NAs:0
```

	Labels	Levels	Class	Storage
age	age		integer	integer
sys.bp	sys bp		integer	integer
sex	sex	2	integer	
junk	junk		integer	integer
state	state		integer	integer

```
+-----+-----+
| Variable | Levels      |
+-----+-----+
|   sex    | female , male |
+-----+-----+
```

```
d <- upData(d,
             state=factor(state, 1:2, c('Alabama','Alaska')),
             rename=c(sys.bp='sbp'),
             labels=c(age = 'Age',
                      sbp = 'Systolic Blood Pressure'),
             drop='junk',      # for > 1: drop=c('junk1','junk2',...)
             units=c(sbp='mmHg'))
```

Input object size:	4192 bytes;	5 variables	4 observations
Renamed variable	sys.bp	to sbp	
Modified variable	state		
Dropped variable	junk		
New object size:	3760 bytes;	4 variables	4 observations

```
contents(d)
```

Data frame:d 4 observations and 4 variables Maximum # NAs:0

	Labels	Units	Levels	Class	Storage
age	Age			integer	integer
sbp	Systolic Blood Pressure	mmHg		integer	integer
sex		sex	2	integer	
state			2	integer	

```
+-----+-----+
| Variable | Levels      |
+-----+-----+
|   sex    | female , male |
+-----+-----+
|   state  | Alabama , Alaska |
+-----+-----+
```

```
describe(d)
```

```
d

4 Variables      4 Observations
-----
age : Age
  n  missing  distinct      Info      Mean      Gmd
  4        0        4        1     36.75     11.83

Value      23    37    42    45
Frequency   1     1     1     1
Proportion 0.25  0.25  0.25  0.25
-----
sbp : Systolic Blood Pressure [mmHg]
  n  missing  distinct      Info      Mean      Gmd
```

```

4          0          4          1      134.8      8.5

Value      127    131    140    141
Frequency   1      1      1      1
Proportion  0.25  0.25  0.25  0.25
-----
sex
  n  missing  distinct
  4        0         2

Value      female    male
Frequency   2        2
Proportion  0.5      0.5
-----
state
  n  missing  distinct
  4        0         2

Value      Alabama  Alaska
Frequency   2        2
Proportion  0.5      0.5
-----
```

```
dim(d); nrow(d); ncol(d); length(d)  # length is no. of variables
```

```
[1] 4 4
```

```
[1] 4
```

```
[1] 4
```

```
[1] 4
```

1.6.4

Defining Small Datasets Inline

For tiny datasets it is easiest to define them as follows:

```
d <- data.frame(age=c(10,20,30), sex=c('male','female','male'),
                 sbp=c(120,125,NA))
```

Large files may be stored in R binary format using `save(..., compress=TRUE)`, which creates an incredibly compact representation of the data in a file usually suffixed with `.rda`. This allows extremely fast loading of the data frame in your next R session using `load(...)`. The `Hmisc` Save and Load functions make this even easier.

1.7

Suggestions for Initial Data Look

The `datadensity` function in the `Hmisc` package gives an overall univariable graphical summary of all variables in the imported dataset. The `contents` and `describe` functions are handy for describing the variables, labels, number of NAs, extreme values, and other values.

1.8

Operating on Data Frames

One of the most common operations is subsetting. In the following example we subset on males older than 26.

```
young.males <- subset(d, sex == 'male' & age > 26)
# If you want to exclude rows that are missing on sex or age:
young.males <- subset(d, sex == 'male' & age > 26 & ! is.na(sex) &
                     ! is.na(age))
# f <- lrm(y ~ sex + age, data=subset(d, sex == 'male' & ...))
# f <- lrm(y ~ sex + age, data=d, subset=sex == 'male' & age > 26 ...)
```

Chapter 2

Algebra Review

2.1

Overview

Algebra and probability are underlying frameworks for basic statistics. The following elements of algebra are particularly important:

- Understanding symbols as variables, and what they can stand for
- Factoring out common terms: $axw + bx = x(aw + b)$
- Factoring out negation of a series of added terms: $-a - b = -(a + b)$
- Simplification of fractions
- Addition, subtraction, multiplication, and division of fractions
- Exponentiation with both fractional and whole number exponents
- Re-writing exponentials of sums: $b^{u+v} = b^u \times b^v$
- Logarithms
 - log to the base b of $x = \log_b x$ is the number y such that $b^y = x$
 - $\log_b b = 1$

- $\log_b b^x = x \log_b b = x$
- $\log_b a^x = x \log_b a$
- $\log_b a^{-x} = -x \log_b a$
- $\log_b(xy) = \log_b x + \log_b y$
- $\log_b \frac{x}{y} = \log_b x - \log_b y$
- When $b = e = 2.71828\dots$, the base of the natural log, $\log_e(x)$ is often written as $\ln x$ or just $\log(x)$
- $\log e = \ln e = 1$
- Anti-logarithms: anti-log to the base b of x is b^x
 - The natural anti-logarithm is e^x , often often written as $\exp(x)$
 - Anti-log is the inverse function of log; it “undoes” a log
- Understanding functions in general, including $\min(x, a)$ and $\max(x, a)$
- Understanding indicator variables such as $[x = 3]$ which can be thought of as true if $x = 3$, false otherwise, or 1 if $x = 3$, 0 otherwise
 - $[x = 3] \times y$ is y if $x = 3$, 0 otherwise
 - $[x = 3] \times [y = 2] = [x = 3 \text{ and } y = 2]$
 - $[x = 3] + 3 \times [y = 2] = 4$ if $x = 3$ and $y = 2$, 3 if $y = 2$ and $x \neq 3$
 - $x \times \max(x, 0) = x^2[x > 0]$
 - $\max(x, 0)$ or $w \times [x > 0]$ are algebraic ways of saying to ignore something if a condition is not met
- Quadratic equations
- Graphing equations

Once you get to multiple regression, some elements of vectors/linear algebra are helpful, for example the vector or dot product, also called the inner product:

- Let x stand for a vector of quantities x_1, x_2, \dots, x_p (e.g., the values of p variables for an animal such as age, blood pressure, etc.)
- Let β stand for another vector of quantities $\beta_1, \beta_2, \dots, \beta_p$ (e.g., weights / regression coefficients / slopes)
- Then $x\beta$ is shorthand for $\beta_1x_1 + \beta_2x_2 + \dots + \beta_px_p$
- $x\beta$ might represent a predicted value in multiple regression, and is known then as the *linear predictor*

2.2

Some Resources

- http://tutorial.math.lamar.edu/pdf/Algebra_Cheat_Sheet.pdf
- <https://www.khanacademy.org/math/algebra>
- <http://biostat.mc.vanderbilt.edu/PrereqAlgebra>
- <http://www.purplemath.com/modules/index.htm>

Chapter 3

General Overview of Biostatistics

There are no routine statistical questions, only questionable statistical routines.

[Sir David R. Cox](#)

It's much easier to get a *result* than it is to get an *answer*.

[Christie Aschwanden,
FireThirtyEight](#)

ABD1.1,p 23-4



3.1

What is Biostatistics?

- Statistics applied to biomedical problems
- Decision making in the face of uncertainty or variability
- Design and analysis of experiments; detective work in observational studies (in epidemiology, outcomes research, etc.)
- Attempt to remove bias or find alternative explanations to those posited by researchers with vested interests
- Experimental design, measurement, description, statistical graphics, data analysis, inference, prediction

To optimize its value, biostatistics needs to be fully integrated into biomedical research and we must recognize that experimental design and execution (e.g., randomization and masking) are all important.

3.1.1

Branches of Statistics

- Frequentist (traditional)
- Bayesian
- Likelihoodist (a bit like Bayes without priors)

See Section 5.3.

3.1.2

Fundamental Principles of Statistics

- Use methods grounded in theory or extensive simulation
- Understand uncertainty
- Design experiments to maximize information and understand sources of variability
- Use all information in data during analysis
- Use discovery and estimation procedures not likely to claim that noise is signal
- Strive for optimal quantification of evidence about effects
- Give decision makers the inputs (*other than the utility function^a*) that optimize decisions
 - Not directly actionable: probabilities that condition on the future to predict the past/present, i.e, those conditioning on the unknown
 - * sensitivity and specificity ($P(\text{test result}|\text{disease status})$)
Sensitivity irrelevant once it is known that the test is +
 - * p -values (condition on effect being zero)
- Present information in ways that are intuitive, maximize information content, and are correctly perceived

^aThe utility function is also called the loss or cost function. It specifies, for example, the damage done by making various decisions such as treating patients who don't have the disease or failing to treat those who do. The optimum Bayes decision is the one that minimizes expected loss. This decision conditions on full information and uses for example predicted risk rather than whether or not the predicted risk is high.

3.2

What Can Statistics Do?

- Refine measurements
- Experimental design
 - Make sure design answers the question
 - Take into account sources of variability
 - Identify sources of bias
 - Developing sequential or adaptive designs
 - Avoid wasting subjects
- (in strong collaboration with epidemiologists) Observational study design
- (in strong collaboration with epidemiologists and philosophers) Causal inference
- Use methods that preserve all relevant information in data
- Robust analysis optimizing power, minimizing assumptions
- Estimating magnitude of effects
- Estimating **shapes** of effects of continuous predictors
- Quantifying causal evidence for effects if the design is appropriate
- Adjusting for confounders
- Properly model effect modification (interaction) / heterogeneity of treatment effect
- Developing and validating predictive models
- Choosing optimum measures of predictive accuracy

- Quantify information added by new measurements / medical tests
- Handling missing data or measurements below detection limits
- Risk-adjusted scorecards (e.g., health provider profiling)
- Visual presentation of results taken into account graphical perception
- Finding alternate explanations for observed phenomena
- Foster reproducible research

See biostat.mc.vanderbilt.edu/BenefitsBasicSci for more benefits of biostatistics.

3.2.1

Statistical Scientific Method

- Statistics is not a bag of tools and math formulas but an evidence-based way of thinking
- It is all important to
 - understand the problem
 - properly frame the question to address it
 - understand and optimize the measurements
 - understand sources of variability
 - much more
- MacKay & Oldford⁶¹ developed a 5-stage representation of the statistical method applied to scientific investigation: **Problem, Plan, Data, Analysis, Conclusion** having the elements below:

Problem	Units & Target Population (Process) Response Variate(s) Explanatory Variates Population Attribute(s) Problem Aspect(s) – causative, descriptive, predictive
Plan	Study Population (Process) (Units, Variates, Attributes) Selecting the response variate(s) Dealing with explanatory variates Sampling Protocol Measuring process Data Collection Protocol
Data	Execute the Plan and record all departures Data Monitoring Data Examination for internal consistency Data storage
Analysis	Data Summary numerical and graphical Model construction build, fit, criticize cycle Formal analysis
Conclusion	Synthesis plain language, effective presentation graphics Limitations of study discussion of potential errors

Recommended Reading for Experimental Design

Glass³⁴, Ruxton and Colegrave⁸⁷, and Chang¹⁵.

Recommended Reading for Clinical Study Design

Hulley *et al.*⁴⁸.

Pointers for Observational Study Design

- Understand the problem and formulate a pertinent question

- Figure out and be able to defend observation periods and “time zero”
- Carefully define subject inclusion/exclusion criteria
- Determine which measurements are required for answering the question while accounting for alternative explanations. Do this **before** examining existing datasets so as to not engage in rationalization bias.
- Collect these measurements or verify that an already existing dataset contains all of them
- Make sure that the measurements are not missing too often and that measurement error is under control. This is even slightly more important for inclusion/exclusion criteria.
- Make sure the use of observational data respects causal pathways. For example don't use outcome/response/late-developing medical complications as if they were independent variables.

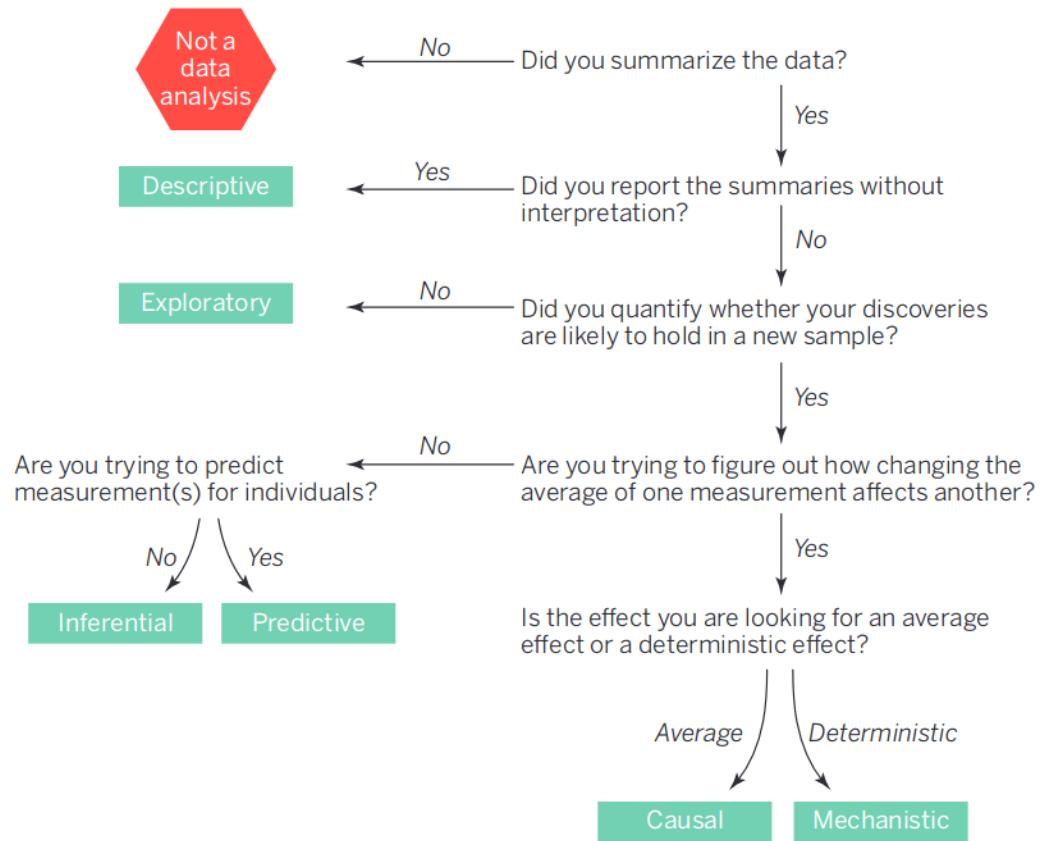
3.3

Types of Data Analysis and Inference

- Description: what happened to *past* patients
- Inference from specific (a sample) to general (a population)
 - Hypothesis testing: test a hypothesis about population or long-run effects
 - Estimation: approximate a population or long term average quantity
- Bayesian inference
 - Data may not be a sample from a population
 - May be impossible to obtain another sample
 - Seeks knowledge of hidden process generating **this** sample (generalization of inference to population)
- Prediction: predict the responses of other patients *like yours* based on analysis of patterns of responses in your patients

Leek and Peng⁵⁹ created a nice data analysis flowchart.

Data analysis flowchart



They also have a succinct summary of common statistical mistakes originating from a failure to match the question with the analysis.

Common mistakes

REAL QUESTION TYPE	PERCEIVED QUESTION TYPE	PHRASE DESCRIBING ERROR
Inferential	Causal	"Correlation does not imply causation"
Exploratory	Inferential	"Data dredging"
Exploratory	Predictive	"Overfitting"
Descriptive	Inferential	"n of 1 analysis"

3.4

Types of Measurements by Their Role in the Study

ABD1.3

- Response variable (clinical endpoint, final lab measurements, etc.)
- Independent variable (predictor or descriptor variable) — something measured when a patient begins to be studied, before the response; often not controllable by investigator, e.g. sex, weight, height, smoking history
- Adjustment variable (confounder) — a variable not of major interest but one needing accounting for because it explains an apparent effect of a variable of major interest or because it describes heterogeneity in severity of risk factors across patients
- Experimental variable, e.g. the treatment or dose to which a patient is randomized; this is an independent variable under the control of the researcher

Table 3.1: Common alternatives for describing independent and response variables

Response variable	Independent variable
Outcome variable	Exposure variable
Dependent variable	Predictor variable
y -variables	x -variable
Case-control group	Risk factor
	Explanatory variable

3.4.1

Proper Response Variables

It is too often the case that researchers concoct response variables Y in such a way that makes the variables *seem* to be easy to interpret, but which contain several hidden problems:

- Y may be a categorization/dichotomization of an underlying continuous response variable. The cutpoint used for the dichotomization is never consistent with data

(see Figure 18.2), is arbitrary (P. 18-11), and causes a huge loss of statistical information and power (P. 18-15).

- Y may be based on a change in a subject's condition whereas what is truly important is the subject's most recent condition (P. 14-11).
- Y may be based on change when the underlying variable is not monotonically related to the ultimate outcome, indicating that positive change is good for some subjects and bad for others (Fig. 14.2).

A proper response variable that optimizes power is one that

- Captures the underlying structure or process
- Has low measurement error
- Has the highest resolution available, e.g.
 - is continuous if the underlying measurement is continuous
 - is ordinal with several categories if the underlying measurement is ordinal
 - is binary only if the underlying process is truly all-or-nothing
- Has the same interpretation for every type of subject, and especially has a direction such that higher values are always good or always bad

3.5

Types of Measurements According to Coding

ABD1.3

- Binary: yes/no, present/absent
- Categorical (aka nominal, polytomous, discrete, multinomial): more than 2 values that are not necessarily in special order
- Ordinal: a categorical variable whose possible values are in a special order, e.g., by severity of symptom or disease; spacing between categories is not assumed to be useful
 - Ordinal variables that are not continuous often have heavy ties at one or more values requiring the use of statistical methods that allow for strange distributions and handle ties well
 - Continuous are also ordinal but ordinal variables may or may not be continuous
- Count: a discrete variable that (in theory) has no upper limit, e.g. the number of ER visits in a day, the number of traffic accidents in a month
- Continuous: a numeric variable having many possible values representing an underlying spectrum
- Continuous variables have the most statistical information (assuming the raw values are used in the data analysis) and are usually the easiest to standardize across hospitals
- Turning continuous variables into categories by using intervals of values is arbitrary and requires more patients to yield the same statistical information (precision or power)
- Errors are not reduced by categorization unless that's the only way to get a subject to answer the question (e.g., income^b)

^bBut note how the Census Bureau tries to maximize the information collected. They first ask for income in dollars. Subjects refusing to answer are asked to choose from among 10 or 20 categories. Those not checking a category are asked to choose from fewer categories.

3.6

Choose Y to Maximize Statistical Information, Power, and Interpretability

bio
log

The outcome (dependent) variable Y should be a high-information measurement that is relevant to the subject at hand. The information provided by an analysis, and statistical power and precision, are strongly influenced by characteristics of Y in addition to the effective sample size.

- Noisy $Y \rightarrow$ variance \uparrow , effect of interest \downarrow
- Low information content/resolution also \rightarrow power \downarrow
- Minimum information Y : binary outcome
- Maximum information Y : continuous response with almost no measurement error
 - Example: measure systolic blood pressure (SBP) well and average 5 readings
- Intermediate: ordinal Y with a few well-populated levels
- Exploration of power vs. number of ordinal Y levels and degree of balance in frequencies of levels: fharrell.com/post/ordinal-info
- See Section 5.12.4 for examples of ordinal outcome scales and interpretation of results

3.6.1

Information Content

- Binary Y : 1 bit
 - all-or-nothing
 - no gray zone, close calls
 - often arbitrary

- SBP: \approx 5 bits
 - range 50-250mmHg (7 bits)
 - accurate to nearest 4mmHg (2 bits)
- Time to binary event: if proportion of subjects having event is small, is effectively a binary endpoint
 - becomes truly continuous and yields high power if proportion with events much greater than $\frac{1}{2}$, if time to event is clinically meaningful
 - if there are multiple events, or you pool events of different severities, time to first event loses information

3.6.2

Dichotomization

Never Dichotomize Continuous or Ordinal Y

- Statistically optimum cutpoint is at the **unknown** population median
 - power loss is still huge
- If you cut at say 2 SDs from the population median, the loss of power can be massive, i.e., may have to increase sample size $\times 4$
- See Sections 18.3.4 and 18.7
- Avoid “responder analysis” (see datamethods.org/t/responder-analysis-loser-x-4)
- Serious ethical issues
- Dumbing-down Y in the quest for clinical interpretability is a mistake. Example:
 - Mean reduction in SBP 7mmHg [2.5, 11.4] for B:A
 - Proportion of pts achieving 10mmHg SBP reduction: A:0.31, B:0.41
 - * Is the difference between 0.31 and 0.41 clinically significant?

- * No information about reductions > 10 mmHg
- Can always restate optimum analysis results in other clinical metrics

3.6.3

Change from Baseline

Never use change from baseline as Y

- Affected by measurement error, regression to the mean
- Assumes
 - you collected a second post-qualification baseline if the variable is part of inclusion/exclusion criteria
 - variable perfectly transformed so that subtraction works
 - post value linearly related to pre
 - slope of pre on post is near 1.0
 - no floor or ceiling effects
 - Y is interval-scaled
- Appropriate analysis (T =treatment)
$$Y = \alpha + \beta_1 \times T + \beta_2 \times Y_0$$
Easy to also allow nonlinear function of Y_0
Also works well for ordinal Y using a semiparametric model
- See Section 14.4 and Chapter 13

3.7

Preprocessing

- In vast majority of situations it is best to analyze the rawest form of the data
- Pre-processing of data (e.g., normalization) is sometimes necessary when the data are high-dimensional
- Otherwise normalizing factors should be part of the final analysis
- A particularly bad practice in animal studies is to subtract or divide by measurements in a control group (or the experimental group at baseline), then to analyze the experimental group as if it is the only group. Many things go wrong:
 - The normalization assumes that there is no biologic variability or measurement error in the control animals' measurements
 - The data may have the property that it is inappropriate to either subtract or divide by other groups' measurements. Division, subtraction, and percent change are highly parametric assumption-laden bases for analysis.
 - A correlation between animals is induced by dividing by a random variable
- A symptom of the problem is a graph in which the experimental group starts off with values 0.0 or 1.0
- The only situation in which pre-analysis normalization is OK in small datasets is in pre-post design or certain crossover studies for which it is appropriate to subtract baseline values from follow-up values

See also Section [4.3.1](#).

3.8

Random Variables



- A potential measurement X
- X might mean a blood pressure that will be measured on a randomly chosen US resident
- Once the subject is chosen and the measurement is made, we have a sample value of this variable
- Statistics often uses X to denote a potentially observed value from some population and x for an already-observed value (i.e., a constant)

But think about the clearer terminology of Richard McElreath^c:

Convention	Proposal
Data	Observed variable
Parameter	Unobserved variable
Likelihood	Distribution
Prior	Distribution
Posterior	Conditional distribution
Estimate	<i>banished</i>
Random	<i>banished</i>

^c<https://youtu.be/yakg94HyWdE?t=2890>

3.9

Probability



- *Probability* traditionally taken as long-run relative frequency
- Example: batting average of a baseball player (long-term proportion of at-bat opportunities resulting in a hit)
- **Not so fast:** The batting average
 - depends on pitcher faced
 - may drop over a season as player tires or is injured
 - drops over years as the player ages
- Getting a hit may be better thought of as a one-time event for which batting average is an approximation of the probability

As described below, the meaning of *probability* is in the mind of the beholder. It can easily be taken to be a long-run relative frequency, a degree of belief, or any metric that is between 0 and 1 that obeys certain basic rules (axioms) such as those of Kolmogorov:

1. A probability is not negative.
2. The probability that at least one of the events in the exhaustive list of possible events occurs is 1.
 - Example: possible events death, nonfatal myocardial infarction (heart attack), or neither
 - $P(\text{at least one of these occurring}) = 1$
3. The probability that at least one of a sequence of mutually exclusive events occurs equals the sum of the individual probabilities of the events occurring.
 - $P(\text{death or nonfatal MI}) = P(\text{death}) + P(\text{nonfatal MI})$

Let A and B denote events, or assertions about which we seek the chances of their veracity. The probabilities that A or B will happen or are true are denoted by $P(A)$, $P(B)$.

The above axioms lead to various useful properties, e.g.

1. A probability cannot be greater than 1.
2. If A is a special case of a more general event or assertion B , i.e., A is a subset of B , $P(A) \leq P(B)$, e.g. $P(\text{animal is human}) \leq P(\text{animal is primate})$.
3. $P(A \cup B)$, the probability of the union of A and B , equals $P(A) + P(B) - P(A \cap B)$ where $A \cap B$ denotes the intersection (joint occurrence) of A and B (the overlap region).
4. If A and B are mutually exclusive, $P(A \cap B) = 0$ so $P(A \cup B) = P(A) + P(B)$.
5. $P(A \cup B) \geq \max(P(A), P(B))$
6. $P(A \cup B) \leq P(A) + P(B)$
7. $P(A \cap B) \leq \min(P(A), P(B))$
8. $P(A|B)$, the conditional probability of A given B holds, is $\frac{P(A \cap B)}{P(B)}$
9. $P(A \cap B) = P(A|B)P(B)$ whether or not A and B are independent. If they are independent, B is irrelevant to $P(A|B)$ so $P(A|B) = P(A)$, leading to the following statement:
10. If a set of events are independent, the probability of their intersection is the product of the individual probabilities.
11. The probability of the union of a set of events (i.e., the probability that at least one of the events occurs) is less than or equal to the sum of the individual event probabilities.
12. The probability of the intersection of a set of events (i.e., the probability that all of the events occur) is less than or equal to the minimum of all the individual probabilities.

So what are examples of what probability might actually mean? In the *frequentist* school, the probability of an event denotes the limit of the long-term fraction of occurrences of the event. This notion of probability implies that the same experiment which generated the outcome of interest can be repeated infinitely often^d.

There are other schools of probability that do not require the notion of replication at all. For example, the school of *subjective* probability (associated with the *Bayesian*

^dBut even a coin will change after 100,000 flips. Likewise, some may argue that a patient is “one of a kind” and that repetitions of the same experiment are not possible. One could reasonably argue that a “repetition” does not denote the same patient at the same stage of the disease, but rather *any* patient with the same *severity* of disease (measured with current technology).

school) “considers probability as a measure of the degree of belief of a given subject in the occurrence of an event or, more generally, in the veracity of a given assertion” (see P. 55 of⁵⁶). de Finetti defined subjective probability in terms of wagers and odds in betting. A risk-neutral individual would be willing to wager $\$P$ that an event will occur when the payoff is \$1 and her subjective probability is P for the event.

As IJ Good has written, the axioms defining the “rules” under which probabilities must operate (e.g., a probability is between 0 and 1) do not define what a probability actually means. He also surmises that all probabilities are subjective, because they depend on the knowledge of the particular observer.

One of the most important probability concepts is that of *conditional probability*. The probability of the veracity of a statement or of an event A occurring given that a specific condition B holds or that an event B has already occurred, is denoted by $P(A|B)$. This is a probability in the presence of knowledge captured by B . For example, if the condition B is that a person is male, the conditional probability is the probability of A for males, i.e., of males, what is the probability of A ? It could be argued that there is no such thing as a completely *unconditional* probability. In this example one is implicitly conditioning on humans even if not considering the person’s sex. Most people would take $P(\text{pregnancy})$ to apply to females.

Conditional probabilities may be computed directly from restricted subsets (e.g., males) or from this formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$. That is, the probability that A is true given B occurred is the probability that both A and B happen (or are true) divided by the probability of the conditioning event B .

Bayes’ rule or theorem is a “conditioning reversal formula” and follows from the basic probability laws: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, read as the probability that event A happens given that event B has happened equals the probability that B happens given that A has happened multiplied by the (unconditional) probability that A happens and divided by the (unconditional) probability that B happens. Bayes’ rule follows immediately from the law of conditional probability, which states that $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

The entire machinery of Bayesian inference derives from only Bayes’ theorem and the basic axioms of probability. In contrast, frequentist inference requires an enormous amount of extra machinery related to the sample space, sufficient statistics, ancillary statistics, large sample theory, and if taking more than one data look, stochastic processes. For many problems we still do not know how to accurately compute a frequentist p -value.

To understand conditional probabilities and Bayes' rule, consider the probability that a randomly chosen U.S. senator is female. As of 2017, this is $\frac{21}{100}$. What is the probability that a randomly chosen female in the U.S. is a U.S. senator?

$$\begin{aligned} P(\text{senator}|\text{female}) &= \frac{P(\text{female}|\text{senator}) \times P(\text{senator})}{P(\text{female})} \\ &= \frac{\frac{21}{100} \times \frac{100}{326M}}{\frac{1}{2}} \\ &= \frac{21}{163M} \end{aligned}$$

So given the marginal proportions of senators and females, we can use Bayes' rule to convert "of senators how many are females" to "of females how many are senators."

The domain of application of probability is all-important. We assume that the true event status (e.g., dead/alive) is unknown, and we also assume that the information the probability is conditional upon (e.g. $P(\text{death}|\text{male, age} = 70)$) is what we would check the probability against. In other words, we do not ask whether $P(\text{death}|\text{male, age} = 70)$ is accurate when compared against $P(\text{death}|\text{male, age}=70, \text{meanbp}=45, \text{patient on downhill course})$. It is difficult to find a probability that is truly not conditional on anything. What is conditioned upon is all important. Probabilities are maximally useful when, as with Bayesian inference, they condition on what is known to provide a forecast for what is unknown. These are "forward time" or "forward information flow" probabilities.

Forward time probabilities can meaningfully be taken out of context more often than backward-time probabilities, as they don't need to consider "what might have happened." In frequentist statistics, the P -value is a backward information flow probability, being conditional on the unknown effect size. This is why P -values must be adjusted for multiple data looks (what might have happened, i.e., what data might have been observed were H_0 true) whereas the current Bayesian posterior probability merely overrides any posterior probabilities computed at earlier data looks, because they condition on current cumulative data.

Chapter 4

Descriptive Statistics, Distributions, and Graphics

4.1

Distributions

The *distribution* of a random variable X is a profile of its variability and other tendencies. Depending on the type of X , a distribution is characterized by the following.

- Binary variable: the probability of “yes” or “present” (for a population) or the proportion of same (for a sample).
- k -Category categorical (polytomous, multinomial) variable: the probability that a randomly chosen person in the population will be from category $i, i = 1, \dots, k$. For a sample, use k proportions or percents.
- Continuous variable: any of the following 4 sets of statistics
 - probability density: value of x is on the x -axis, and the relative likelihood of observing a value “close” to x is on the y -axis. For a sample this yields a histogram.
 - cumulative probability distribution: the y -axis contains the probability of observing $X \leq x$. This is a function that is always rising or staying flat, never

decreasing. For a sample it corresponds to a cumulative histogram^a

- all of the *quantiles* or *percentiles* of X
 - all of the *moments* of X (mean, variance, skewness, kurtosis, . . .)
 - If the distribution is characterized by one of the above four sets of numbers, the other three sets can be derived from this set
- Ordinal Random Variables
 - Because there may be heavy ties, quantiles may not be good summary statistics
 - The mean may be useful if the spacings have reasonable quantitative meaning
 - The mean is especially useful for summarizing ordinal variables that are counts
 - When the number of categories is small, simple proportions may be helpful
 - With a higher number of categories, exceedance probabilities or the empirical cumulative distribution function are very useful
 - Knowing the distribution we can make intelligent guesses about future observations from the same series, although unless the distribution really consists of a single point there is a lot of uncertainty in predicting an individual new patient's response. It is less difficult to predict the average response of a group of patients once the distribution is known.
 - At the least, a distribution tells you what proportion of patients you would expect to see whose measurement falls in a given interval.

4.1.1

Continuous Distributions

```
x ← seq(-3, 35, length=150)
par(mfrow=c(1,2)); xl ← expression(x)    # Fig. 4.1:
plot(x, dt(x, 4, 6), type='l', xlab=xl, ylab='Probability Density Function')
plot(x, pt(x, 4, 6), type='l', xlab=xl, ylab='Cumulative Distribution Function')
```

^aBut this *empirical cumulative distribution function* can be drawn with no grouping of the data, unlike an ordinary histogram.

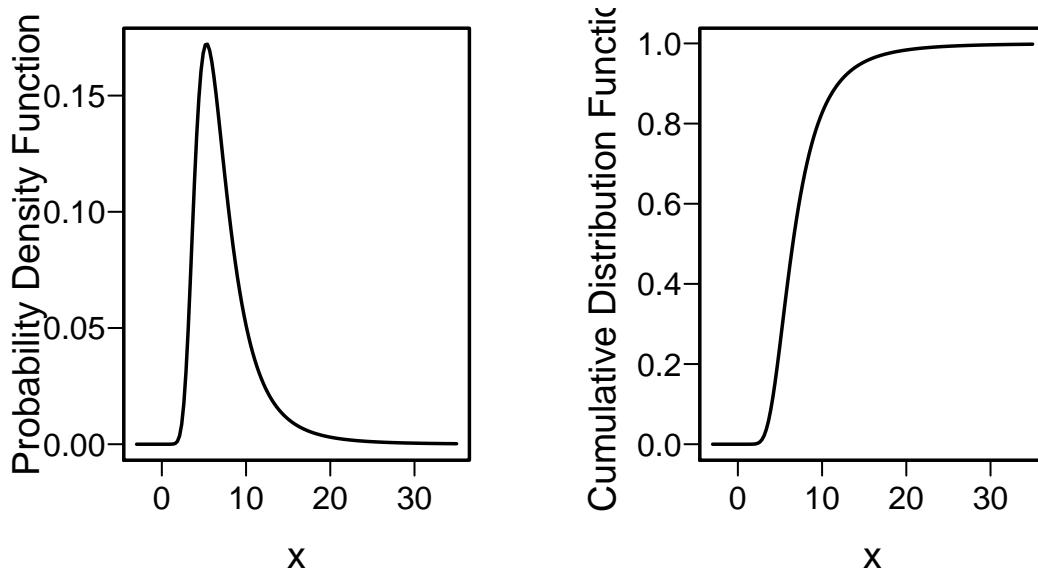


Figure 4.1: Example probability density (a) and cumulative probability distribution (b) for a positively skewed random variable (skewed to the right)

```
set.seed(1); x ← rnorm(1000)      # Fig. 4.2:
hist(x, nclass=40, prob=TRUE, col=gray(.9), xlab=xl, ylab='')
x ← seq(-4, 4, length=150)
lines(x, dnorm(x, 0, 1), col='blue', lwd=2)
```

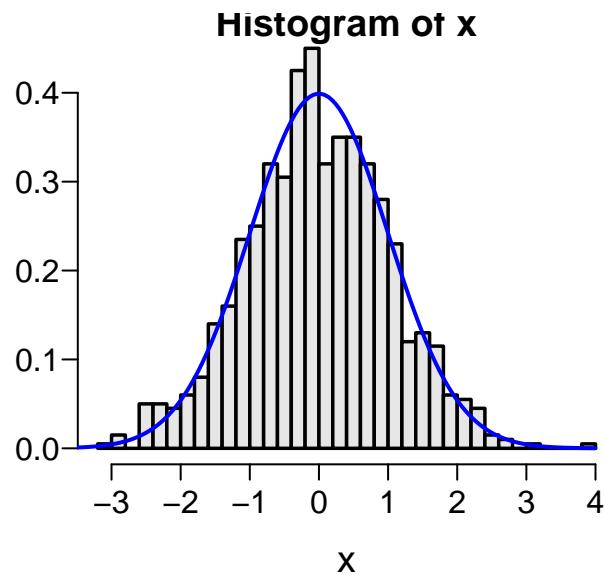


Figure 4.2: Example of a continuous distribution that is symmetric: the Gaussian (normal) distribution with mean 0 and variance 1, along with a histogram from a sample of size 1000 from this distribution

```
set.seed(2)
x ← c(rnorm(500, mean=0, sd=1), rnorm(500, mean=6, sd=3))
hist(x, nclass=40, prob=TRUE, col=gray(.9), xlab=xl, ylab='')
lines(density(x), col='blue', lwd=2)
abline(v=c(0, 6), col='red')    # Fig. 4.3
```

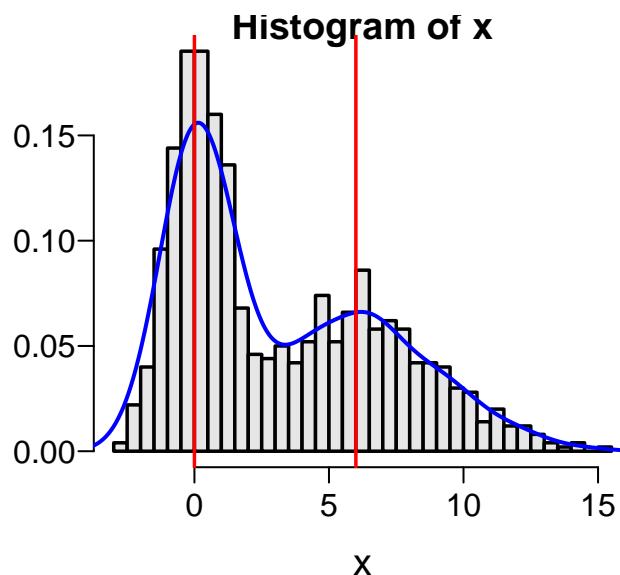


Figure 4.3: Example of a bimodal distribution from sampling from a mixture of normal distributions with different means and variances and estimating the underlying density function. Vertical red lines indicate true population means of the two component populations. Such a distribution can occur naturally or by failing to condition on a binary characteristic such as sex.

4.1.2

Ordinal Variables

- Continuous ratio-scaled variables are ordinal
- Not all ordinal variables are continuous or ratio-scaled
- Best to analyze ordinal response variables using nonparametric tests or ordinal regression
- Heavy ties may be present
- Often better to treat count data as ordinal rather than to assume a distribution such as Poisson or negative binomial that is designed for counts
 - Poisson or negative binomial do not handle extreme clumping at zero
- Example ordinal variables are below

```
x <- 0:14
y <- c(.8, .04, .03, .02, rep(.01, 11))
plot(x, y, xlab="x", ylab='', type='n')    # Fig. 4.4
segments(x, 0, x, y)
```

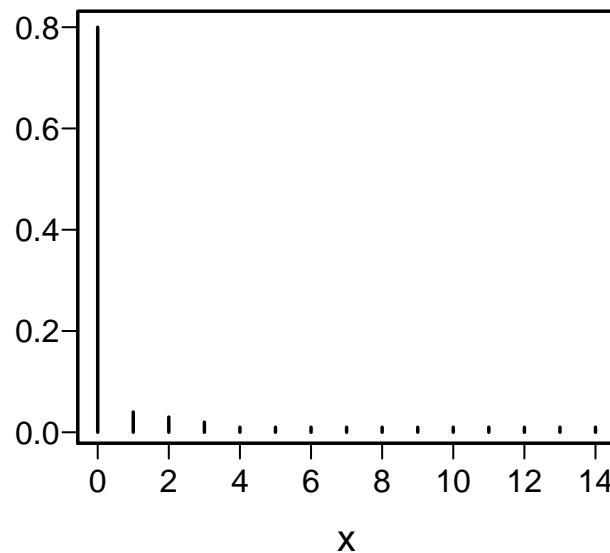


Figure 4.4: Distribution of number of days in the hospital in the year following diagnosis

```
x <- 1:10
y <- c(.1, .13, .18, .19, 0, 0, .14, .12, .08, .06)
plot(x, y, xlab=xl, ylab='', type='n')      # Fig. 4.5
segments(x, 0, x, y)
```

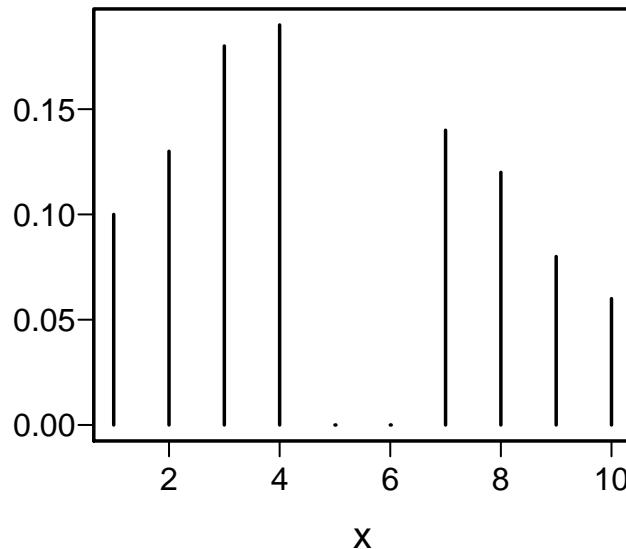


Figure 4.5: Distribution of a functional status score that does not have points in the middle

The `getHdata` function in the `Hmisc` package⁴⁰ finds, downloads, and `load()`s datasets from biostat.mc.vanderbilt.edu/DataSets.

```
require(Hmisc)
getHdata(nhgh)    # NHANES dataset      Fig. 4.6:
scr <- pmin(nhgh$SCr, 5)  # truncate at 5 for illustration
scr[scr == 5 | runif(nrow(nhgh)) < .05] <- 5  # pretend 1/20 dialyzed
hist(scr, nclass=50, xlab='Serum Creatinine', ylab='Density', prob=TRUE)
```

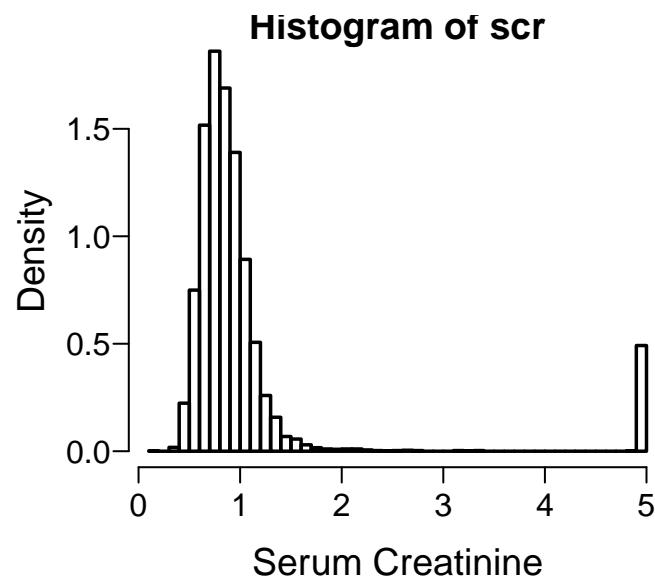


Figure 4.6: Distribution of serum creatinine where the patient requiring dialysis is taken to have the worst renal function. The variable is mostly continuous but is best analyzed as ordinal so that no assumption is made about how to score dialysis except for being worse than all non-dialysis patients. Data taken from NHANES where no patients were actually dialyzed.

4.2

Descriptive Statistics

4.2.1

Categorical Variables

ABD3

- Proportions of observations in each category

Note: The mean of a binary variable coded 1/0 is the proportion of ones.

- For variables representing counts (e.g., number of comorbidities), the mean is a good summary measure (but not the median)
- Modal (most frequent) category

4.2.2

Continuous Variables

Denote the sample values as x_1, x_2, \dots, x_n

Measures of Location

“Center” of a sample

- Mean: arithmetic average

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

Population mean μ is the long-run average (let $n \rightarrow \infty$ in computing \bar{x})

- center of mass of the data (balancing point)
- highly influenced by extreme values even if they are highly atypical

- Median: middle sorted value, i.e., value such that $\frac{1}{2}$ of the values are below it and above it
 - always descriptive
 - unaffected by extreme values
 - not a good measure of central tendency when there are heavy ties in the data
 - if there are heavy ties and the distribution is limited or well-behaved, the mean often performs better than the median (e.g., mean number of diseased fingers)
- Geometric mean: hard to interpret and affected by low outliers; better to use median

Quantiles

Quantiles are general statistics that can be used to describe central tendency, spread, symmetry, heavy tailedness, and other quantities.

- Sample median: the 0.5 quantile or 50th percentile
- Quartiles Q_1, Q_2, Q_3 : 0.25 0.5 0.75 quantiles or 25th, 50th, 75th percentiles
- Quintiles: by 0.2
- In general the p th sample quantile x_p is the value such that a fraction p of the observations fall below that value
- p th population quantile: value x such that the probability that $X \leq x$ is p

Spread or Variability

- Interquartile range: Q_1 to Q_3
Interval containing $\frac{1}{2}$ of the subjects

Meaningful for any continuous distribution

- Other quantile intervals
- Variance (for symmetric distributions): averaged squared difference between a randomly chosen observation and the mean of all observations

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

The -1 is there to increase our estimate to compensate for our estimating the center of mass from the data instead of knowing the population mean.^b

- Standard deviation: $s = \sqrt{\text{variance}}$
 - $\sqrt{\text{average squared difference of an observation from the mean}}$
 - can be defined in terms of proportion of sample population within ± 1 SD of the mean **if the population is normal**
- SD and variance are not useful for very asymmetric data, e.g. “the mean hospital cost was \$10000 \pm \$15000”
- Gini’s mean difference: mean absolute difference over all possible pairs of observations. This is highly interpretable, robust, and useful for all interval-scaled data, and is even highly precise if the data are normal^c.
- range: not recommended because range \uparrow as $n \uparrow$ and is dominated by a single outlier
- coefficient of variation: not recommended (depends too much on how close the mean is to zero)

Example of Gini’s mean difference for describing patient-to-patient variability of systolic blood pressure: If Gini’s mean difference is 7mmHg, this means that the average disagreement (absolute difference) between any two patients is 7mmHg.

^b \bar{x} is the value of μ such that the sum of squared values about μ is a minimum.

^cGini’s mean difference is labeled `Gmd` in the output of the R `Hmisc describe` function, and may be computed separately using the `Hmisc GiniMd` function

4.3

Graphics

Cleveland^{19,18} is the best source of how-to information on making scientific graphs. Much information may be found at <http://biostat.mc.vanderbilt.edu/StatGraphCourse>, especially these notes: <http://goo.gl/DHE0a2>. Information about graphics for reporting clinical trials may be found at biostat.mc.vanderbilt.edu/RCTGraphics. A link to John Rauser's exceptional video about principles of good graphics is found on that page as well as in the movie icon in the right margin.



Murrell⁷¹ has an excellent summary of recommendations:

- Display data values using position or length.
- Use horizontal lengths in preference to vertical lengths.
- Watch your data–ink ratio.
- Think very carefully before using color to represent data values.
- Do not use areas to represent data values.
- *Please* do not use angles or slopes to represent data values.
- *Please, please* do not use volumes to represent data values.

On the fifth point above, avoid the use of *bars* when representing a single number. Bar widths contain no information and get in the way of important information. This is addressed below.

R has superior graphics implemented in multiple models, including

- Base graphics such as `plot()`, `hist()`, `lines()`, `points()` which give the user maximum control and are best used when not stratifying by additional variables other than the ones being summarized
- The `lattice` package which is fast but not quite as good as `ggplot2` when one needs to vary more than one of color, symbol, size, or line type due to having more than

one categorizing variable

- The `ggplot2` package which is very flexible and has the nicest defaults especially for constructing keys (legends/guides)
- For semi-interactive graphics inserted into html reports, the R `plotly` package, which uses the `plotly` system (which uses the Javascript `d3` library) is extremely powerful. See <https://plot.ly/r/getting-started>.
- Fully interactive graphics can be built using `RShiny` but this requires a server to be running while the graph is viewed.

For `ggplot2`, <http://www.cookbook-r.com/Graphs> contains a nice cookbook. See also <http://learnr.wordpress.com>. To get excellent documentation with examples for any `ggplot2` function, google `ggplot2 functionname`. `ggplot2` graphs can be converted into `plotly` graphics using the `ggplotly` function. But you will have more control using R `plotly` directly.

The older non-interactive graphics models which are useful for producing printed and `pdf` output are starting to be superceded with interactive graphics. One of the biggest advantages of the latter is the ability to present the most important graphic information front-and-center but to allow the user to easily hover the mouse over areas in the graphic to see tabular details.

4.3.1

Graphing Change vs. Raw Data

A common mistake in scientific graphics is to cover up subject variability by normalizing repeated measures for baseline (see Section 3.7). Among other problems, this prevents the reader from seeing regression to the mean for subjects starting out at very low or very high values, and from seeing variation in intervention effect as a function of baseline values. It is highly recommended that all the raw data be shown, including those from time zero. When the sample size is not huge, spaghetti plots are most effective for graphing longitudinal data because all points from the same subject over time are connected. An example [23, pp. 161-163] is below.

```
require(Hmisc)    # also loads ggplot2
getHdata(cdystonia)
ggplot(cdystonia, aes(x=week, y=twstrs, color=factor(id))) +
```

```
geom_line() + xlab('Week') + ylab('TWSTRS-total score') +
facet_grid(treat ~ site) +
guides(color=FALSE) # Fig. 4.7
```

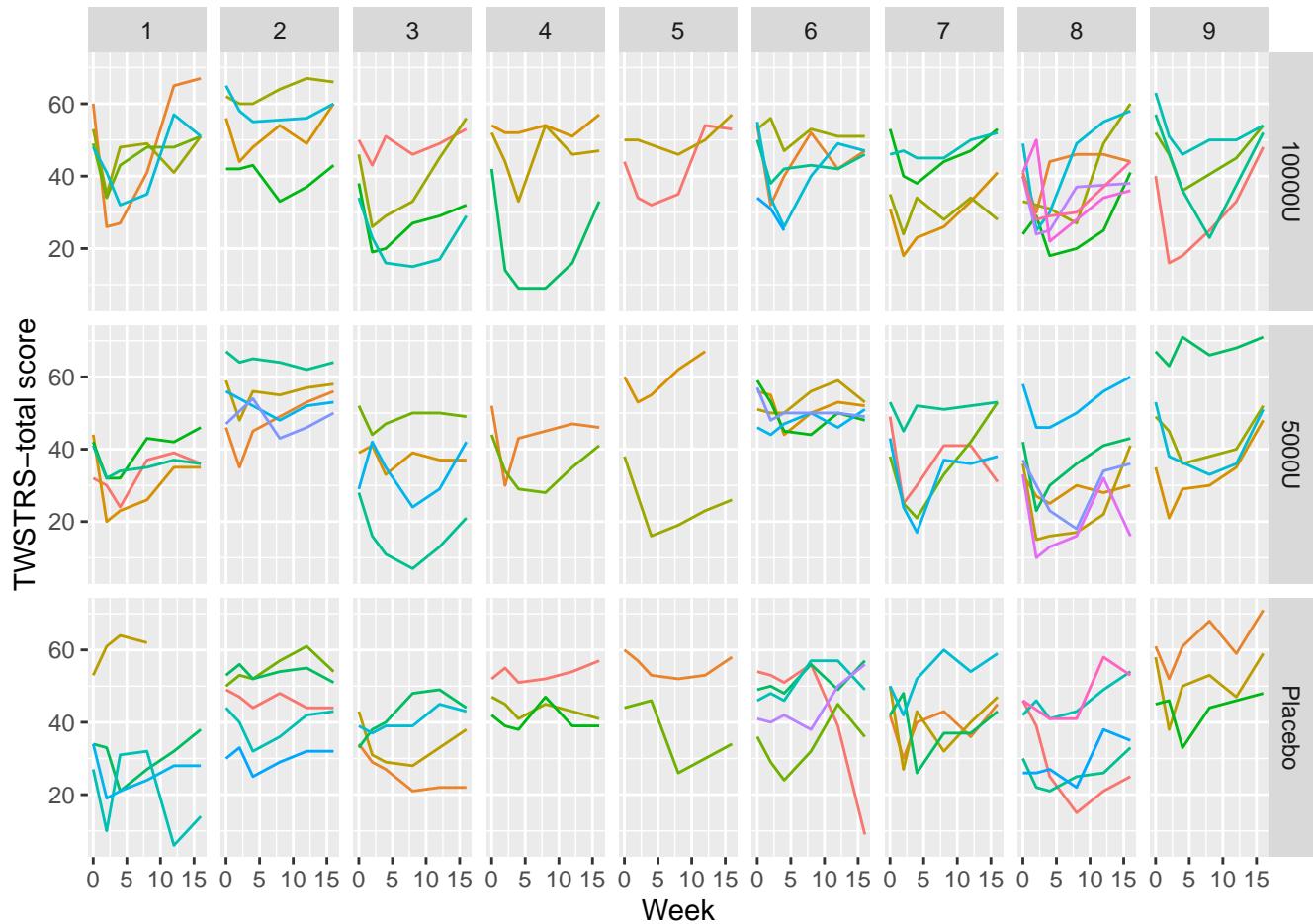


Figure 4.7: Spaghetti plot showing all the raw data on the response variable for each subject, stratified by dose and study site (1–9). Importantly, week 0 (baseline) measurements are included.

Graphing the raw data is usually essential.

4.3.2

Categorical Variables

- pie chart
 - high ink:information ratio
 - optical illusions (perceived area or angle depends on orientation vs. horizon)
 - hard to label categories when many in number

- bar chart

- high ink:information ratio
- hard to depict confidence intervals (one sided error bars?)
- hard to interpret if use subcategories
- labels hard to read if bars are vertical

- dot chart

- leads to accurate perception
- easy to show all labels; no caption needed
- allows for multiple levels of categorization (see Figures 4.8 and 4.9)

```
getHdata(titanic3)
d <- upData(titanic3,
            agec      = cut2(age, c(10, 15, 20, 30)), print=FALSE)
d <- with(d, as.data.frame(table(sex, pclass, agec)))
d <- subset(d, Freq > 0)
ggplot(d, aes(x=Freq, y=agec)) + geom_point() +
  facet_grid(sex ~ pclass) +
  xlab('Frequency') + ylab('Age')
```

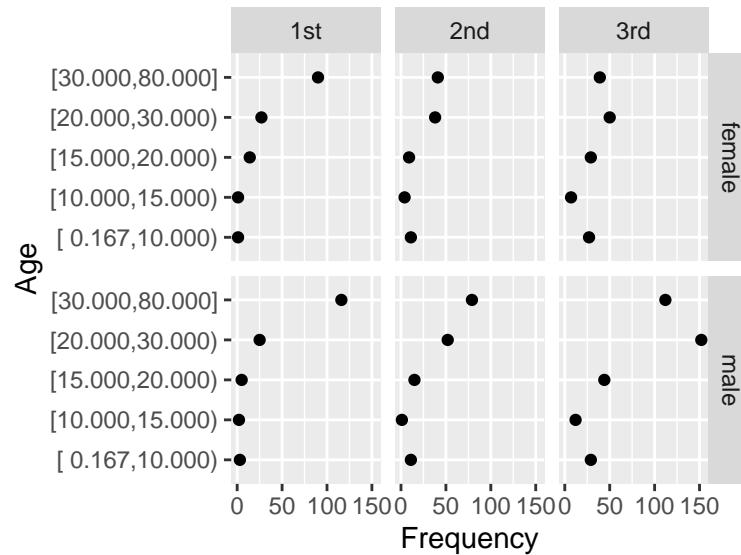


Figure 4.8: Dot chart showing frequencies from cross-classifications of discrete variables for Titanic passengers

- * multi-panel display for multiple major categorizations
- * lines of dots arranged vertically within panel

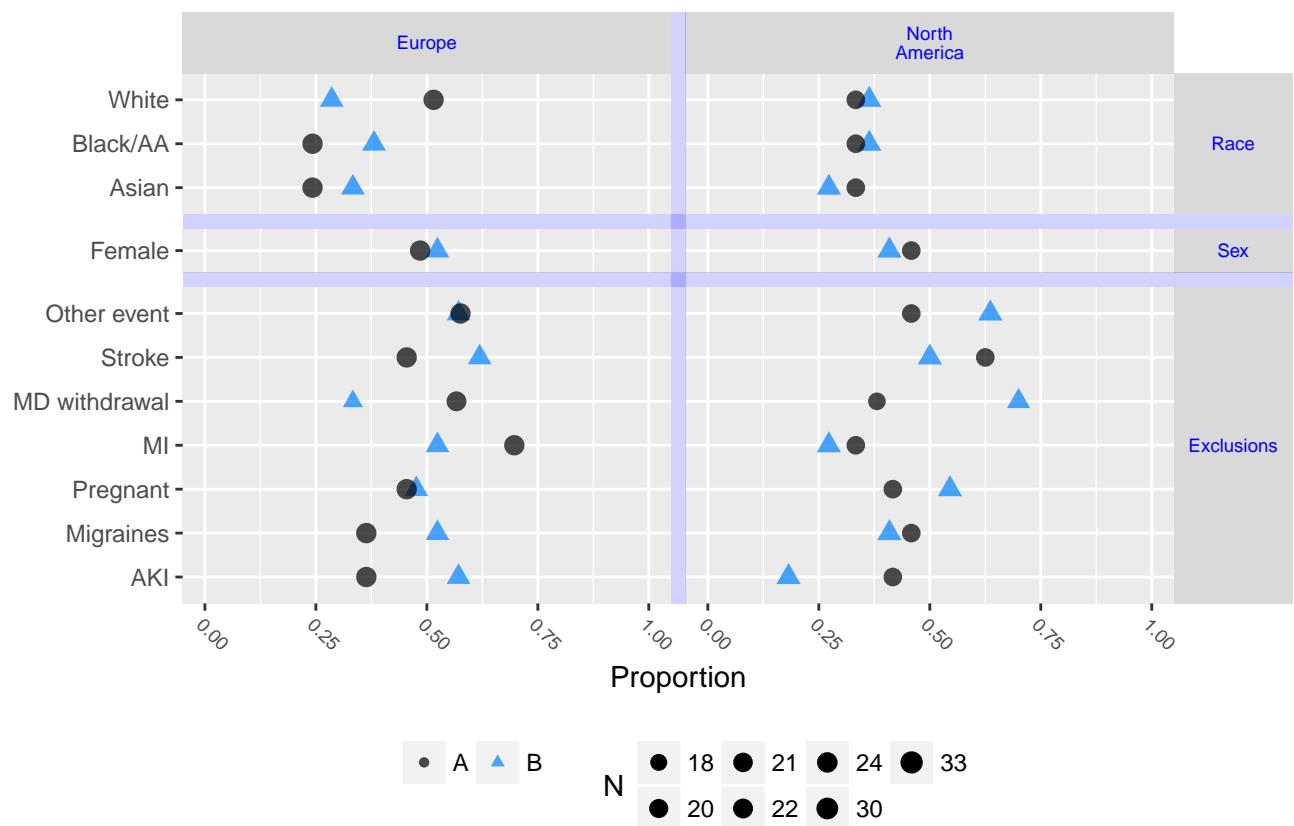


Figure 4.9: Dot chart for categorical demographic variables, stratified by treatment and region

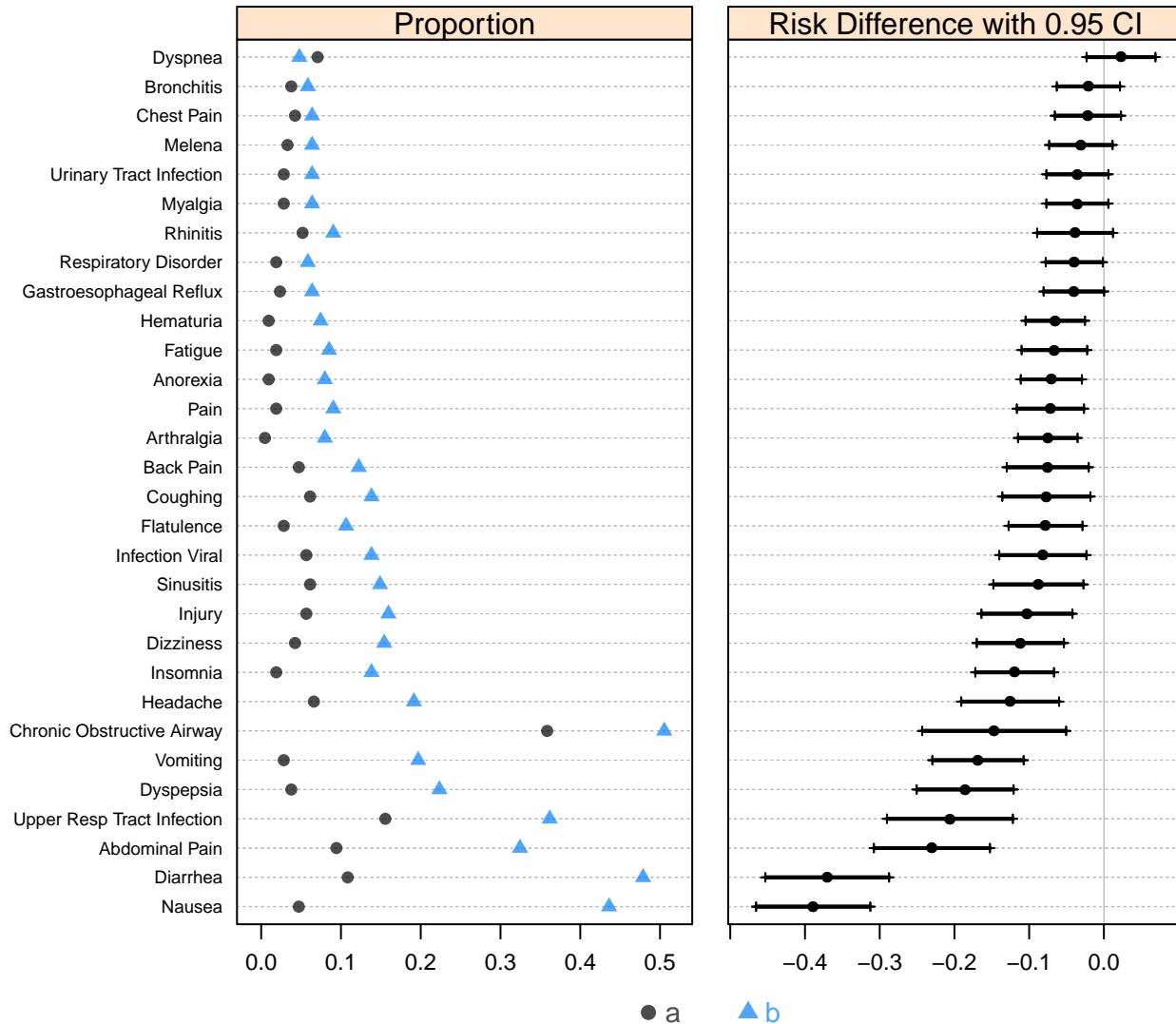


Figure 4.10: Dot chart showing proportion of subjects having adverse events by treatment, sorted by risk difference, produced by the R greport package. See test.Rnw at biostat.mc.vanderbilt.edu/Report

- * categories within a single line of dots
- easy to show 2-sided error bars
- Avoid chartjunk such as dummy dimensions in bar charts, rotated pie charts, use of solid areas when a line suffices

4.3.3

Continuous Variables

Raw Data

For graphing two continuous variable, scatterplots are often essential. The following example draws a scatterplot on a very large number of observations in a measurement comparison study where the goal is to measure esophageal pH longitudinally and across subjects.

```
getHdata(esoph)
contents(esoph)
```

```
Data frame:esoph          136127 observations and 2 variables      Maximum # NAs:0
```

	Labels	Class	Storage
orophar	Esophageal pH by Oropharyngeal Device	numeric	double
conv	Esophageal pH by Conventional Device	numeric	double

```
xl <- label(esoph$conv)
yl <- label(esoph$orophar)
ggplot(esoph, aes(x=conv, y=orophar)) + geom_point(pch='.') +
  xlab(xl) + ylab(yl) +    # Fig. 4.11
  geom_abline(intercept = 0, slope = 1)
```

With large sample sizes there are many collisions of data points and hexagonal binning is an effective substitute for the raw data scatterplot. The number of points represented by one hexagonal symbol is stratified into 20 groups of approximately equal numbers of points. The code below is not currently working for the ggplot2 package version 2.1.0.

```
ggplot(esoph, aes(x=conv, y=orophar)) +
  stat_binhex(aes(alpha=..count.., color=Hmisc::cut2(..count.., g=20)),
              bins=80) +
  xlab(xl) + ylab(yl) +
  guides(alpha=FALSE, fill=FALSE, color=guide_legend(title='Frequency'))
```

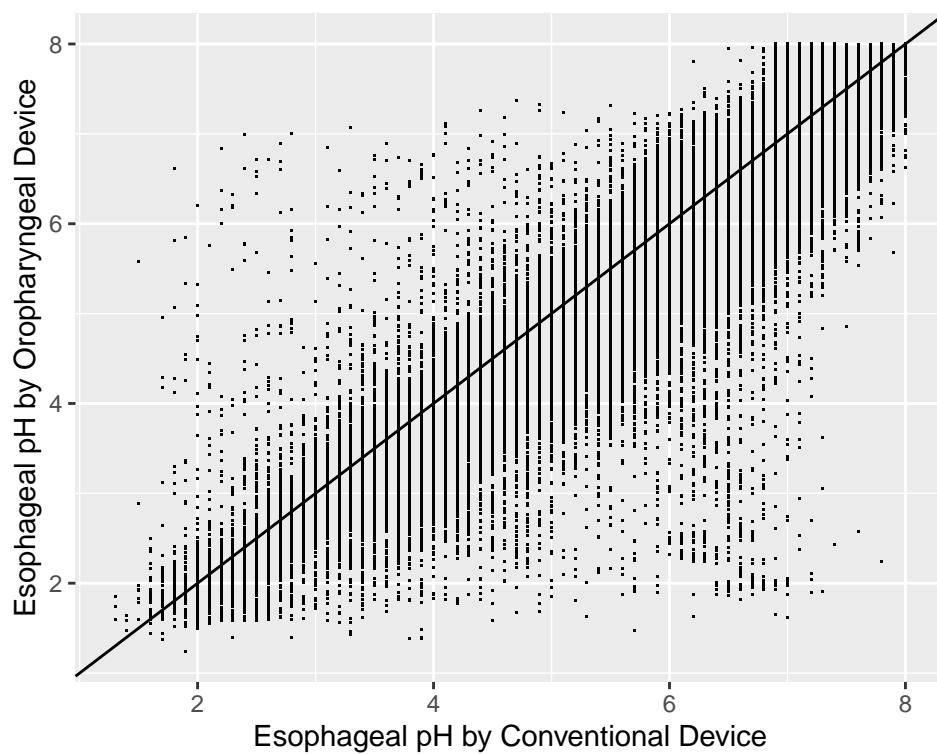


Figure 4.11: Scatterplot of one measurement mode against another

Instead we use the `Hmisc ggfreqScatter` function to bin the points and represent frequencies of overlapping points with color and transparency levels.

```
with(esopH, ggfreqScatter(conv, orophar, bins=50, g=20) +
     geom_abline(intercept=0, slope=1)) # Fig. 4.12
```

Distributions

- histogram showing relative frequencies
 - requires arbitrary binning of data
 - not optimal for comparing multiple distributions
- cumulative distribution function: proportion of values $\leq x$ vs. x (Figure 4.13)
Can read all quantiles directly off graph.

```
getHdata(pbc)
pbcr <- subset(pbc, drug != 'not randomized')
Ecdf(pbcr[,c('bili','albumin','protime','sgot')]), # Fig. 4.13
      group=pbcr$drug, col=1:2,
      label.curves=list(keys='lines'))
```

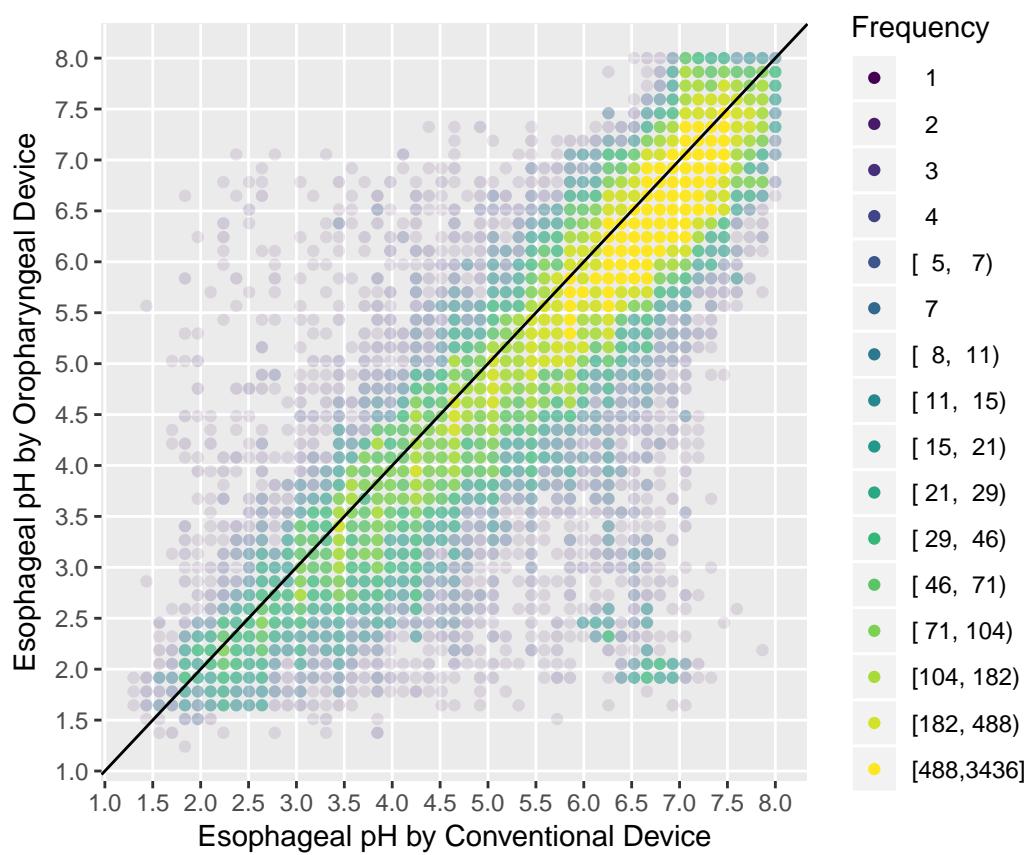


Figure 4.12: Binned points (2500 total bins) with frequency counts shown as color and transparency level

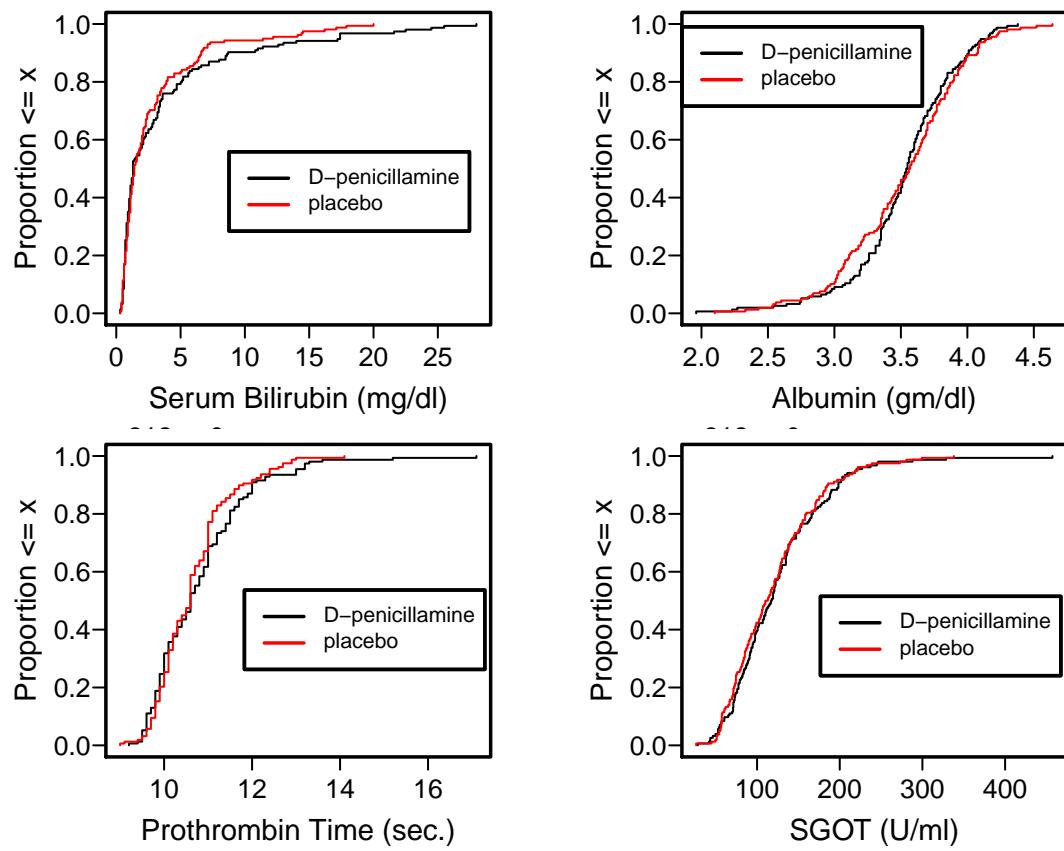


Figure 4.13: Empirical cumulative distributions of baseline variables stratified by treatment in a randomized controlled trial. m is the number of missing values.

- Box plots shows quartiles plus the mean. They are a good way to compare many groups as seen in Figures 4.14 and 4.16.

```
getHdata(support) # Fig. 4.14
bwplot(dzgroup ~ crea, data=support, panel=panel.bpplot,
       probs=c(.05,.25), xlim=c(0,8), xlab='Serum Creatinine')
```

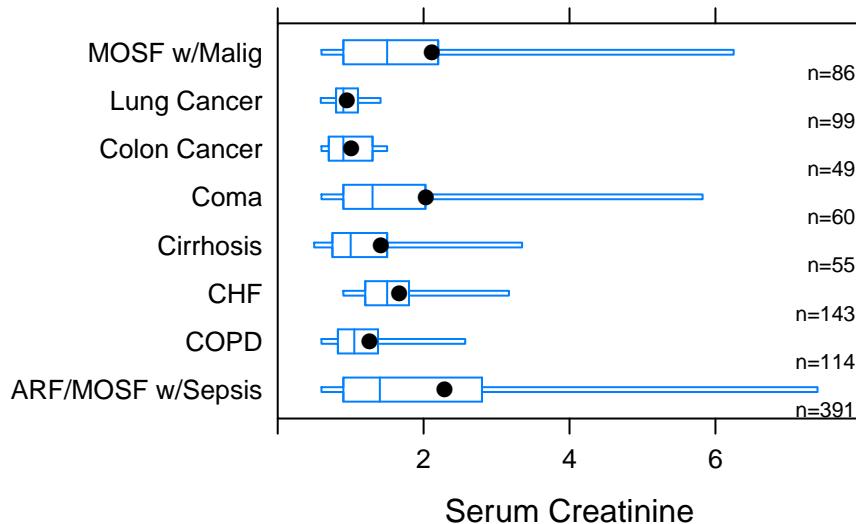


Figure 4.14: Box plots showing the distribution of serum creatinine stratified by major diagnosis. Dot: mean; vertical line: median; large box:interquartile range. The 0.05 and 0.95 quantiles are also shown, which is not the way typical box plots are drawn but is perhaps more useful. Asymmetry of distributions can be seen by both disagreement between $Q_3 - Q_2$ and $Q_2 - Q_1$ and by disagreement between Q_2 and \bar{x} .

Figure 4.16 uses extended box plots. The following schematic shows how to interpret them.

```
bpplt() # Fig. 4.15
```

```
require(lattice) # Fig. 4.16:
getHdata(diabetes)
wst ← cut2(diabetes$waist, g=2)
levels(wst) ← paste('Waist', levels(wst))
bwplot(cut2(age,g=4) ~ glyhb | wst*gender, data=diabetes,
       panel=panel.bpplot, xlab='Glycosylated Hemoglobin', ylab='Age Quartile')
```

Box plots are inadequate for displaying bimodality. Violin plots show the entire distribution well if the variable being summarized is fairly continuous.

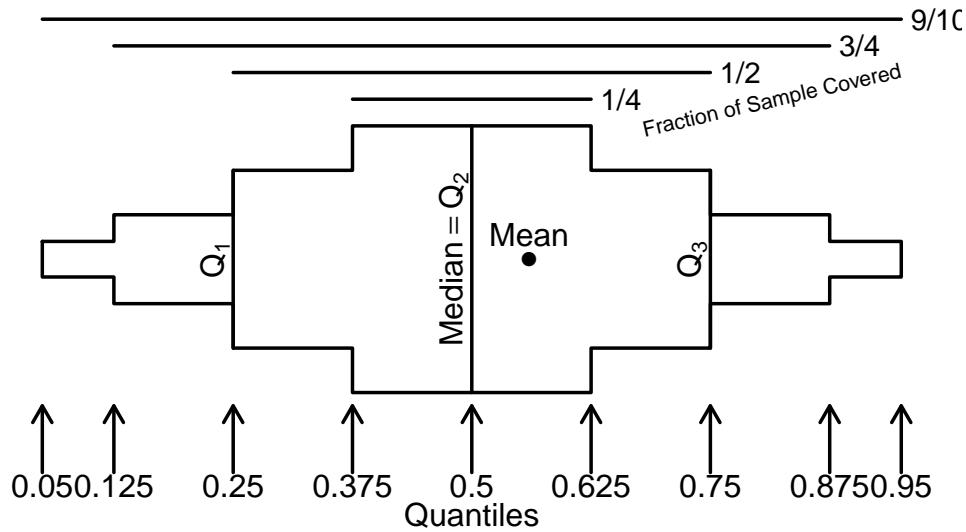


Figure 4.15: Schematic for extended box plot

Relationships

- When response variable is continuous and descriptor (stratification) variables are categorical, multi-panel dot charts, box plots, multiple cumulative distributions, etc., are useful.
- Two continuous variables: scatterplot

4.3.4

Graphs for Summarizing Results of Studies

- Dot charts with optional error bars (for confidence limits) can display any summary statistic (proportion, mean, median, mean difference, etc.)
- It is not well known that the confidence interval for a difference in two means cannot be derived from individual confidence limits.^d
Show individual confidence limits as well as actual CLs for the difference.

^dIn addition, it is not necessary for two confidence intervals to be separated for the difference in means to be significantly different from zero.

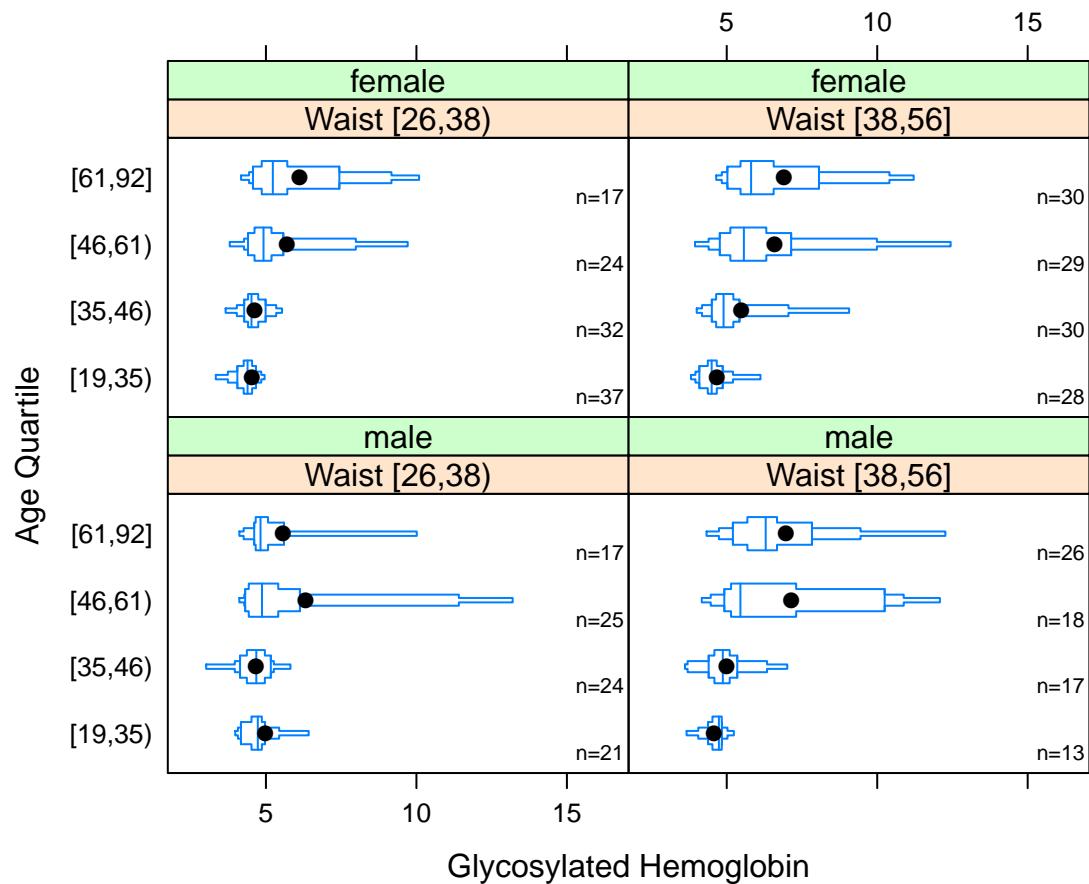


Figure 4.16: Extended box plots for glycohemoglobin stratified by quartiles of age (vertical), two-tiles of waist circumference (horizontal), and sex (vertical)

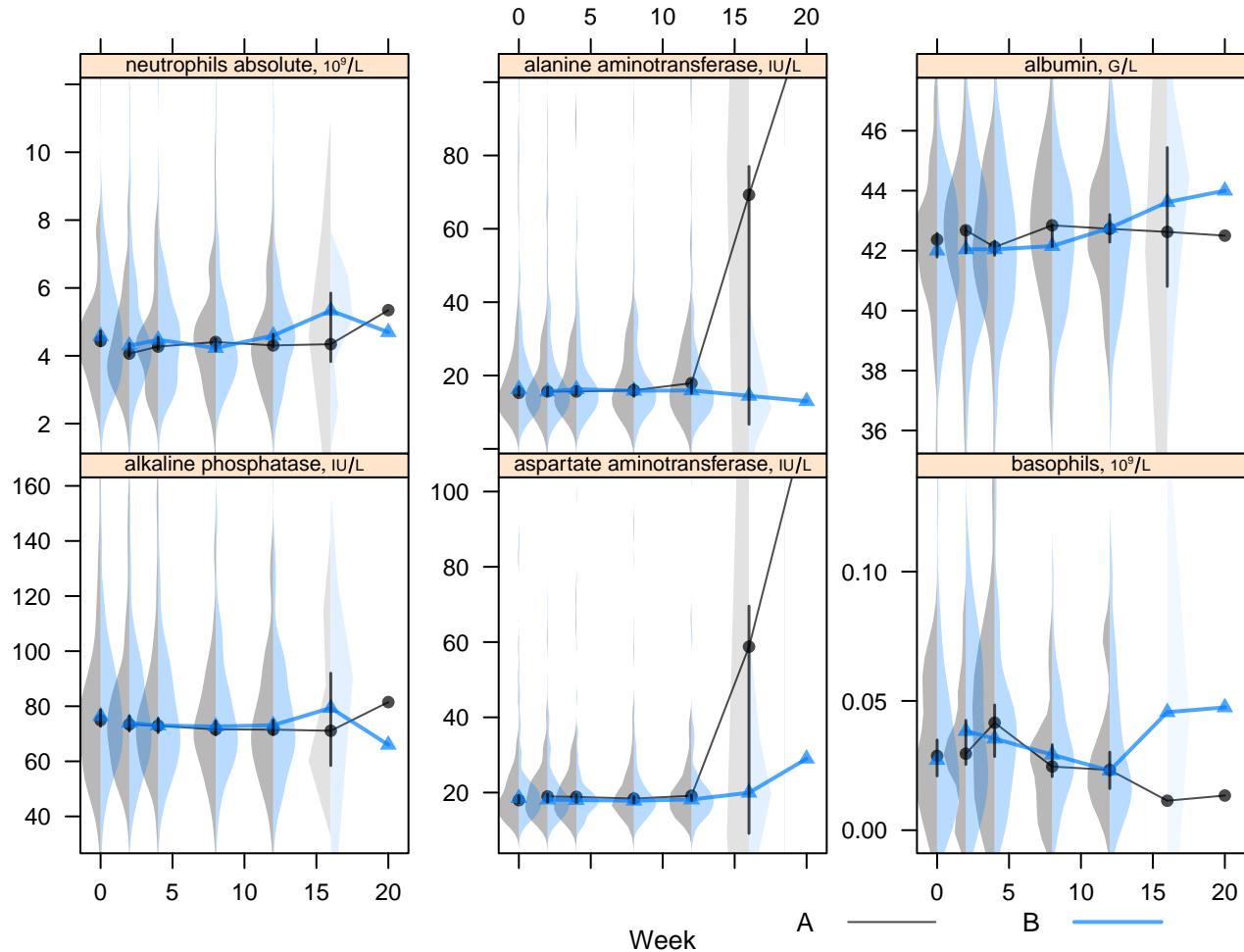


Figure 4.17: One-half violin plots for longitudinal data, stratified by treatment. Density estimates for groups with insufficient sample sizes are faded. Density plots are back-to-back for treatment A and B. Points are treatment medians. When the black vertical line does not touch the two medians, the medians are significantly different at the $\alpha = 0.05$ level. Graphic was produced by the R `greport` package.

```

attach(diabetes)
set.seed(1)
male    ← smean.cl.boot(glyhb[gender=='male'],   reps=TRUE)
female ← smean.cl.boot(glyhb[gender=='female'], reps=TRUE)
dif    ← c(mean=male['Mean']-female['Mean'],
          quantile(attr(male, 'reps')-attr(female, 'reps'), c(.025,.975)))
plot(0,0,xlab='Glycated Hemoglobin',ylab='',   # Fig. 4.18
     xlim=c(5,6.5),ylim=c(0,4), axes=F)
axis(1, at=seq(5, 6.5, by=0.25))
axis(2, at=c(1,2,3.5), labels=c('Female','Male','Difference'),
     las=1, adj=1, lwd=0)
points(c(male[1],female[1]), 2:1)
segments(female[2], 1, female[3], 1)
segments(male[2], 2, male[3], 2)
offset ← mean(c(male[1],female[1])) - dif[1]
points(dif[1] + offset, 3.5)
segments(dif[2]+offset, 3.5, dif[3]+offset, 3.5)
at ← c(-.5,-.25,0,.25,.5,.75,1)
axis(3, at=at+offset, label=format(at))
segments(offset, 3, offset, 4.25, col=gray(.85))
abline(h=c(2 + 3.5)/2, col=gray(.85))

```

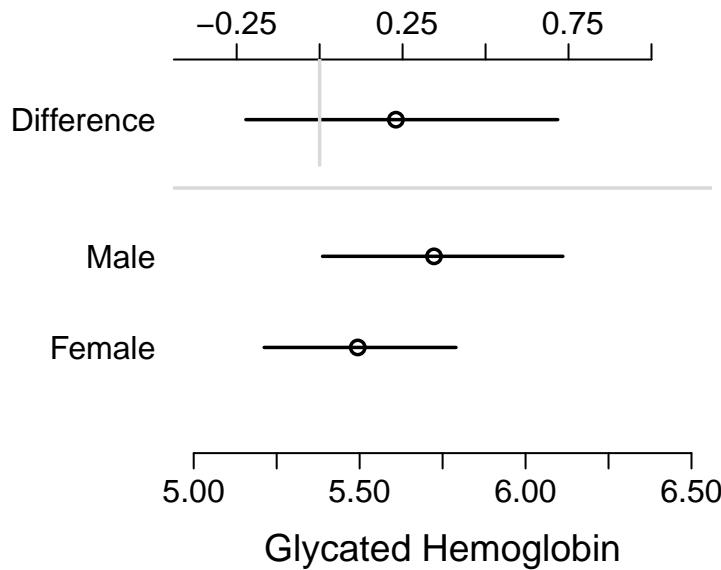


Figure 4.18: Means and nonparametric bootstrap 0.95 confidence limits for glycated hemoglobin for males and females, and confidence limits for males - females. Lower and upper x -axis scales have same spacings but different centers. Confidence intervals for differences are generally wider than those for the individual constituent variables.

- For showing relationship between two continuous variables, a trend line or regression model fit, with confidence bands

Bar Plots with Error Bars

- “Dynamite” Plots
- Height of bar indicates mean, lines represent standard error
- High ink:information ratio
- Hide the raw data, assume symmetric confidence intervals
- Replace with
 - Dot plot (smaller sample sizes)
 - Box plot (larger sample size)

```
getHdata(FEV); set.seed(13)
FEV <- subset(FEV, runif(nrow(FEV)) < 1/8)    # 1/8 sample
require(ggplot2)
s <- with(FEV, summarize(fev, llist(sex, smoke), smean.cl.normal))
ggplot(s, aes(x=smoke, y=fev, fill=sex)) +      # Fig. 4.19
  geom_bar(position=position_dodge(), stat="identity") +
  geom_errorbar(aes(ymin=Lower, ymax=Upper),
                 width=.1,
                 position=position_dodge(.9))
```

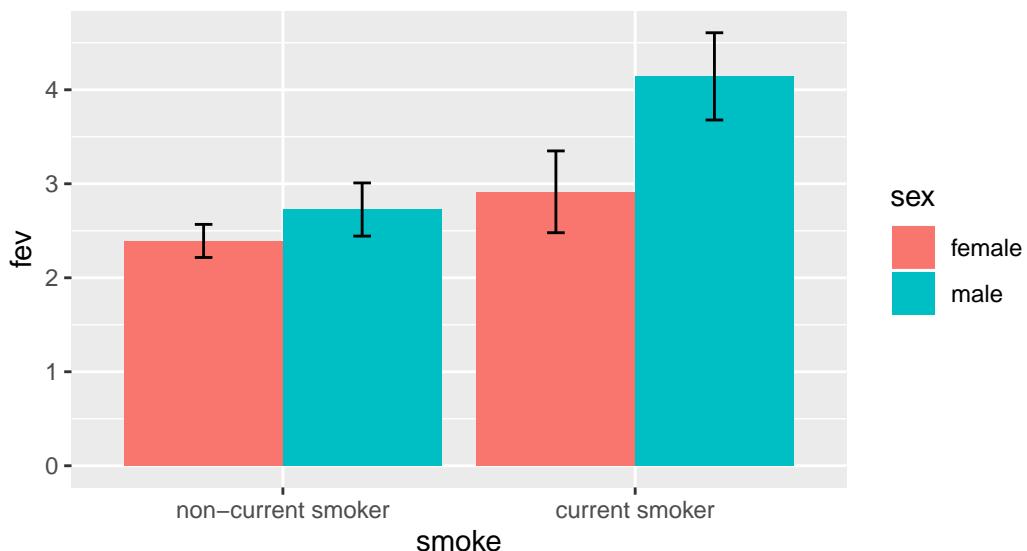


Figure 4.19: Bar plot with error bars—“dynamite plot”

See <http://biostat.mc.vanderbilt.edu/DynamitePlots> for a list of the many problems caused by dynamite plots, plus some solutions.

Instead of the limited information shown in the bar chart, show the raw data along with box plots. Modify default box plots to replace whiskers with the interval between 0.1 and 0.9 quantiles.

```
require(ggplot2)    # Fig. 4.20
stats <- function(x) {
  z <- quantile(x, probs=c(.1, .25, .5, .75, .9))
  names(z) <- c('ymin', 'lower', 'middle', 'upper', 'ymax')
  if(length(x) < 10) z[c(1,5)] <- NA
  z
}
ggplot(FEV, aes(x=sex, y=fev)) +
  stat_summary(fun.data=stats, geom='boxplot', aes(width=.75), shape=5,
              position='dodge', col='lightblue') +
  geom_dotplot(binaxis='y', stackdir='center', position='dodge', alpha=.4) +
  stat_summary(fun.y=mean, geom='point', shape=5, size=4, color='blue') +
  facet_grid(~ smoke) +
  xlab('') + ylab(expression(FEV[1])) + coord_flip()
```

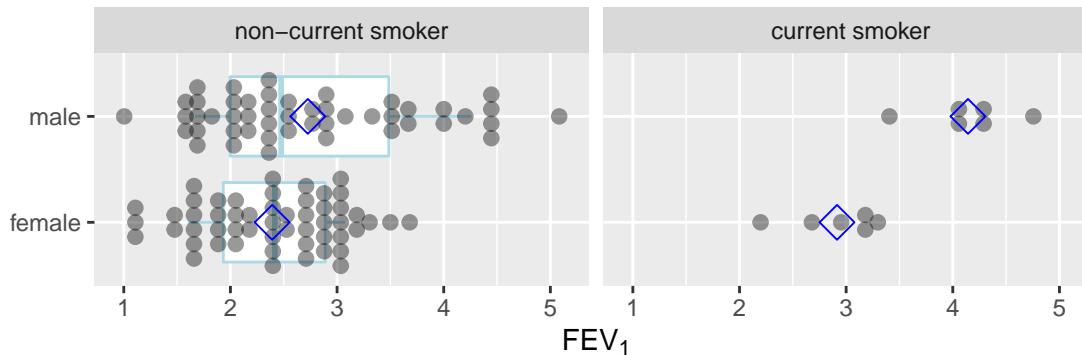


Figure 4.20: Jittered raw data and box plots. Middle vertical lines indicate medians and diamonds indicate means. Horizontal lines indicate 0.1 to 0.9 quantiles when $n \geq 10$. The ink:information ratio for this plot is far better than a dynamite plot.

Use a violin plot to show the distribution density estimate (and its mirror image) instead of a box plot.

```
ggplot(FEV, aes(x=sex, y=fev)) +
  geom_violin(width=.6, col='lightblue') +
  geom_dotplot(binaxis='y', stackdir='center', position='dodge', alpha=.4) +
  stat_summary(fun.y=median, geom='point', color='blue', shape='+', size=12) +
  facet_grid(~ smoke) +
  xlab('') + ylab(expression(FEV[1])) + coord_flip()
```

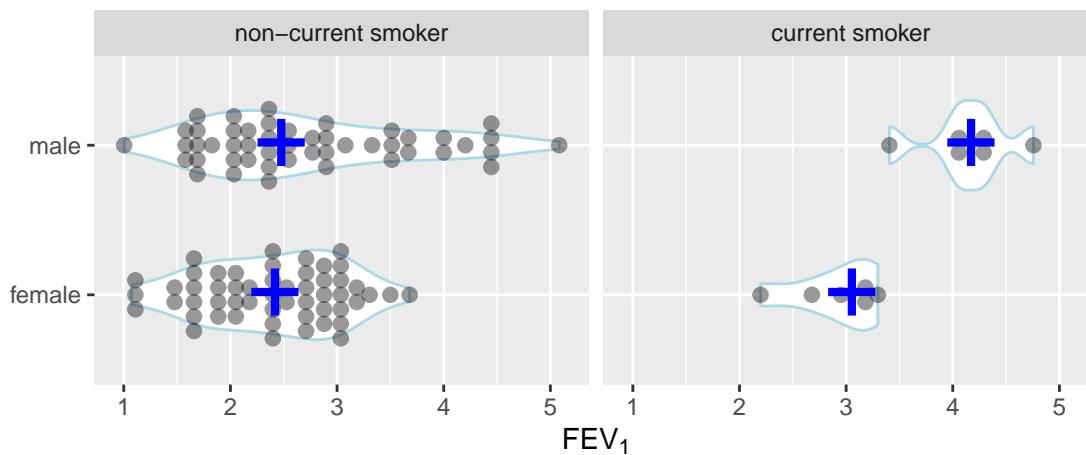


Figure 4.21: Jittered raw data and violin plots with median indicated by +

Semi-Interactive Graphics Examples

These examples are all found in hbiostat.org/talks/rmedicine19.html.

- R for Clinical Trial Reporting
- Spike histograms with hovertext for overall statistical summary (slide 11)
- Dot plots (slide 12)
- Extended box plots (slide 13)
- Spike histograms with quantiles, mean, dispersion (slide 15)
- Survival plots with CI for difference, continuous number at risk (slide 16)
- Example clinical trial reports (slide 35)

Other examples: descriptions of BBR course participants: hbiostat.org/bbr/registrants.html.

4.3.5

Graphs for Describing Statistical Model Fits

Several types of graphics are useful. These are all implemented in the R `rms` package⁴¹.

Partial effect plots : Show the effect on Y of varying one predictor at a time, holding the other predictors to medians or modes, and include confidence bands. This is the best approach for showing shapes of effects of continuous predictors.

Effect charts : Show the difference in means, odds ratio, hazard ratio, fold change, etc., varying each predictor and holding others to medians or modes^e. For continuous variables that do not operate linearly, this kind of display is not very satisfactory because of its strong dependence on the settings over which the predictor is set. By default inter-quartile-range effects are used.

Nomograms : Shows the entire model if the number of interactions is limited. Nomograms show strengths and shapes of relationships, are very effective for continuous predictors, and allow computation of predicted values (although without confidence limits).

Here are examples using NHANES data to predict glycohemoglobin from age, sex, race/ethnicity, and BMI.

Note: ordinary regression is not an adequate fit for glycohemoglobin; an excellent fit comes from ordinal regression. BMI is not an adequate summary of body size. The following ordinary regression model in the -1.75 power of glycohemoglobin resulted in approximately normal residuals and is used for illustration. The transformation is subtracted from a constant just so that positive regression coefficients indicate that increasing a predictor increases glycohemoglobin. The inverse transformation raises predicted values to the $-\frac{1}{1.75}$ power after accounting for the subtraction, and is used to estimate the median glycohemoglobin on the original scale^f. Restricted cubic spline functions with 4 default knots are used to allow age and BMI to act smoothly but nonlinearly. Partial effects plots are in Fig. 4.22.

```
require(rms)
getHdata(nhgh)    # NHANES data
dd <- datadist(nhgh); options(datadist='dd')
g      <- function(x) 0.09 - x ^ - (1 / 1.75)
ginverse <- function(y) (0.09 - y) ^ -1.75
f <- ols(g(gh) ~ rcs(age, 4) + re + sex + rcs(bmi, 4), data=nhgh)
```

^eIt does not matter what the other variables are set to if they do not interact with the variable being varied.

^fIf residuals have a normal distribution after transforming the dependent variable, the estimated mean and median transformed values are the same. Inverse transforming the estimates provides an estimate of the median on the original scale (but not the mean).

```
cat('{\small\n')
```

```
f
```

Linear Regression Model

```
ols(formula = g(gh) ~ rcs(age, 4) + re + sex + rcs(bmi, 4), data = nhgh)
```

	Model Likelihood Ratio Test		Discrimination Indexes	
Obs 6795	LR χ^2	1861.16	R^2	0.240
σ 0.0235	d.f.	11	R^2_{adj}	0.238
d.f. 6783	Pr(> χ^2)	0.0000	g	0.015

	Residuals				
	Min	1Q	Median	3Q	Max
	-0.09736	-0.01208	-0.002201	0.008237	0.1689

	$\hat{\beta}$	S.E.	t	Pr(> t)
Intercept	-0.2884	0.0048	-60.45	<0.0001
age	0.0002	0.0001	3.34	0.0008
age'	0.0010	0.0001	7.63	<0.0001
age"	-0.0040	0.0005	-8.33	<0.0001
re=Other Hispanic	-0.0013	0.0011	-1.20	0.2318
re=Non-Hispanic White	-0.0082	0.0008	-10.55	<0.0001
re=Non-Hispanic Black	-0.0013	0.0009	-1.34	0.1797
re=Other Race Including Multi-Racial	0.0006	0.0014	0.47	0.6411
sex=female	-0.0022	0.0006	-3.90	<0.0001
bmi	-0.0006	0.0002	-2.54	0.0111
bmi'	0.0059	0.0009	6.44	<0.0001
bmi"	-0.0161	0.0025	-6.40	<0.0001

```
print(anova(f), dec.ss=3, dec.ms=3)
```

Analysis of Variance for g(gh)

	d.f.	Partial SS	MS	F	P
age	3	0.732	0.244	441.36	<0.0001
Nonlinear	2	0.040	0.020	35.83	<0.0001
re	4	0.096	0.024	43.22	<0.0001
sex	1	0.008	0.008	15.17	<0.0001
bmi	3	0.184	0.061	110.79	<0.0001
Nonlinear	2	0.023	0.011	20.75	<0.0001
TOTAL NONLINEAR	4	0.068	0.017	30.94	<0.0001
REGRESSION	11	1.181	0.107	194.29	<0.0001
ERROR	6783	3.749	0.001		

```
cat('}\n')
```

```
# Show partial effects of all variables in the model, on the original scale
ggplot(Predict(f, fun=ginv), # Fig. 4.22
       ylab=expression(paste('Predicted Median ', HbA['1c'])))
```

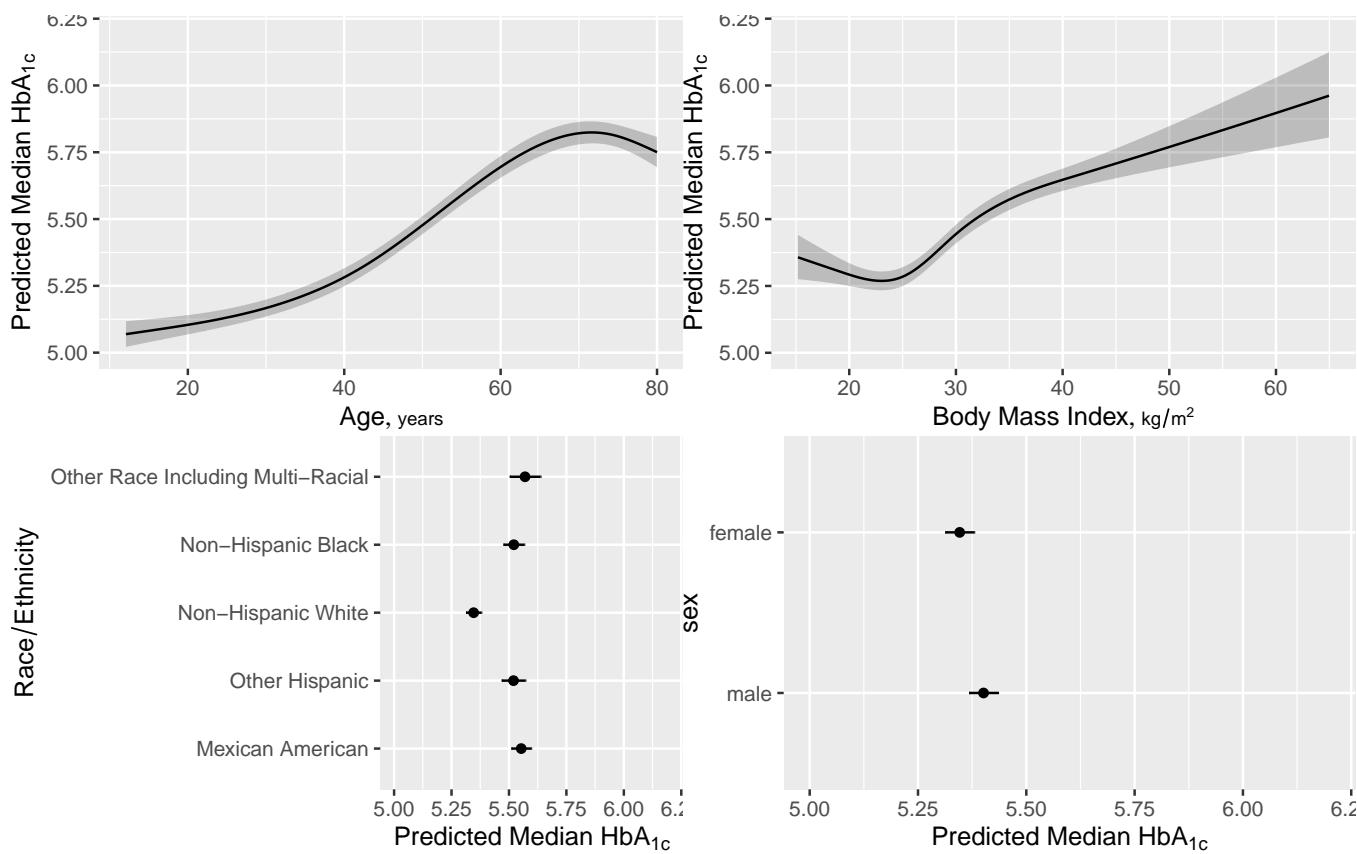


Figure 4.22: Partial effects in NHANES HbA1c model

An effect chart is in Fig. 4.23 and a nomogram is in Fig. 4.24. See <http://stats.stackexchange.com/questions/155430/clarifications-regarding-reading-a-nomogram> for excellent examples showing how to read such nomograms.

```
plot(summary(f))    # Fig. 4.23
```

```
plot(nomogram(f, fun=ginverse, funlabel='Median HbA1c'))    # Fig. 4.24
```

Graphing Effect of Two Continuous Variables on Y

The following examples show the estimated combined effects of two continuous predictors on outcome. The two models included interaction terms, the second example using penalized maximum likelihood estimation with a tensor spline in diastolic \times systolic blood pressure.

Figure 4.26 is particularly interesting because the literature had suggested (based on approximately 24 strokes) that pulse pressure was the main cause of hemorrhagic stroke whereas this flexible modeling approach (based on approximately 230 strokes) suggests

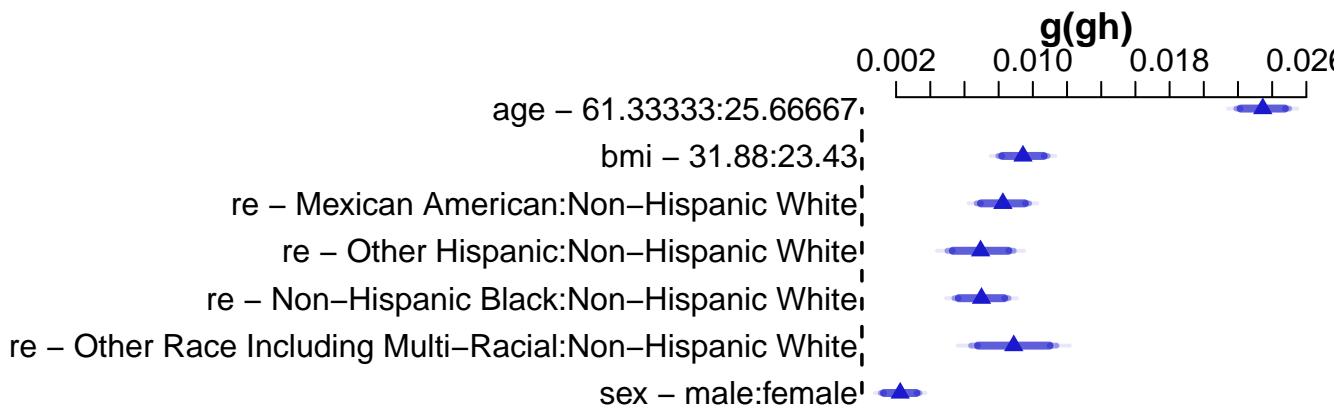


Figure 4.23: Partial effects chart on the transformed scale. For age and BMI, effects are inter-quartile-range effects. 0.9, 0.95, and 0.99 confidence limits are shown.

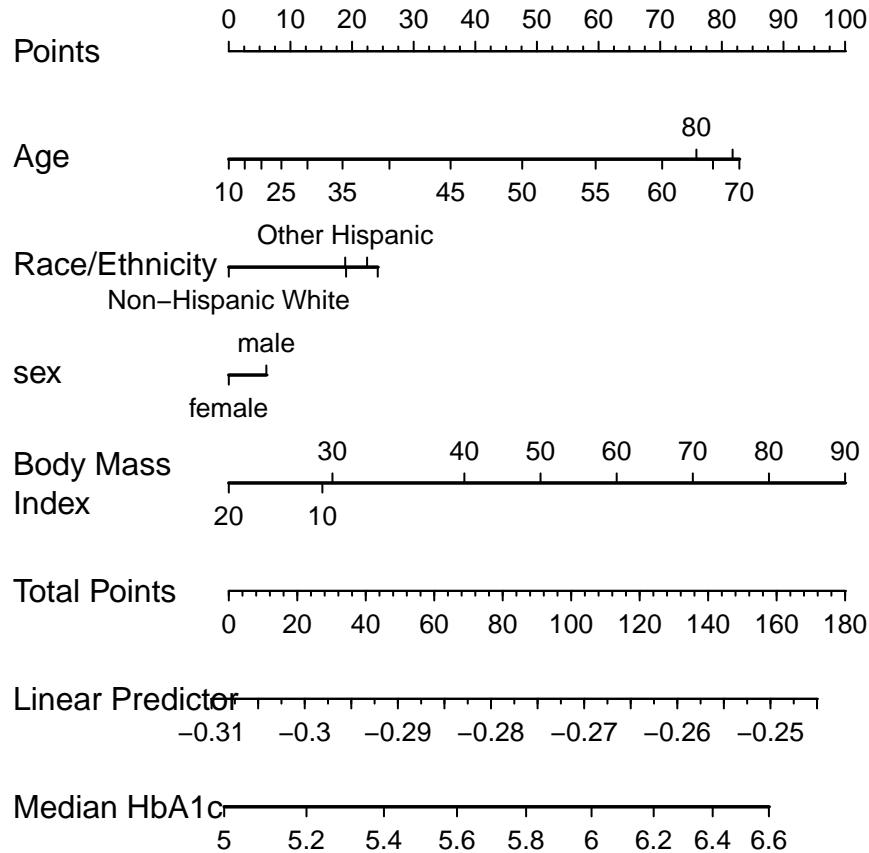


Figure 4.24: Nomogram for predicting median HbA_{1c}. To use the nomogram, use the top Points scale to convert each predictor value to a common scale. Add the points and read this number on the Total Points scale, then read down to the median.

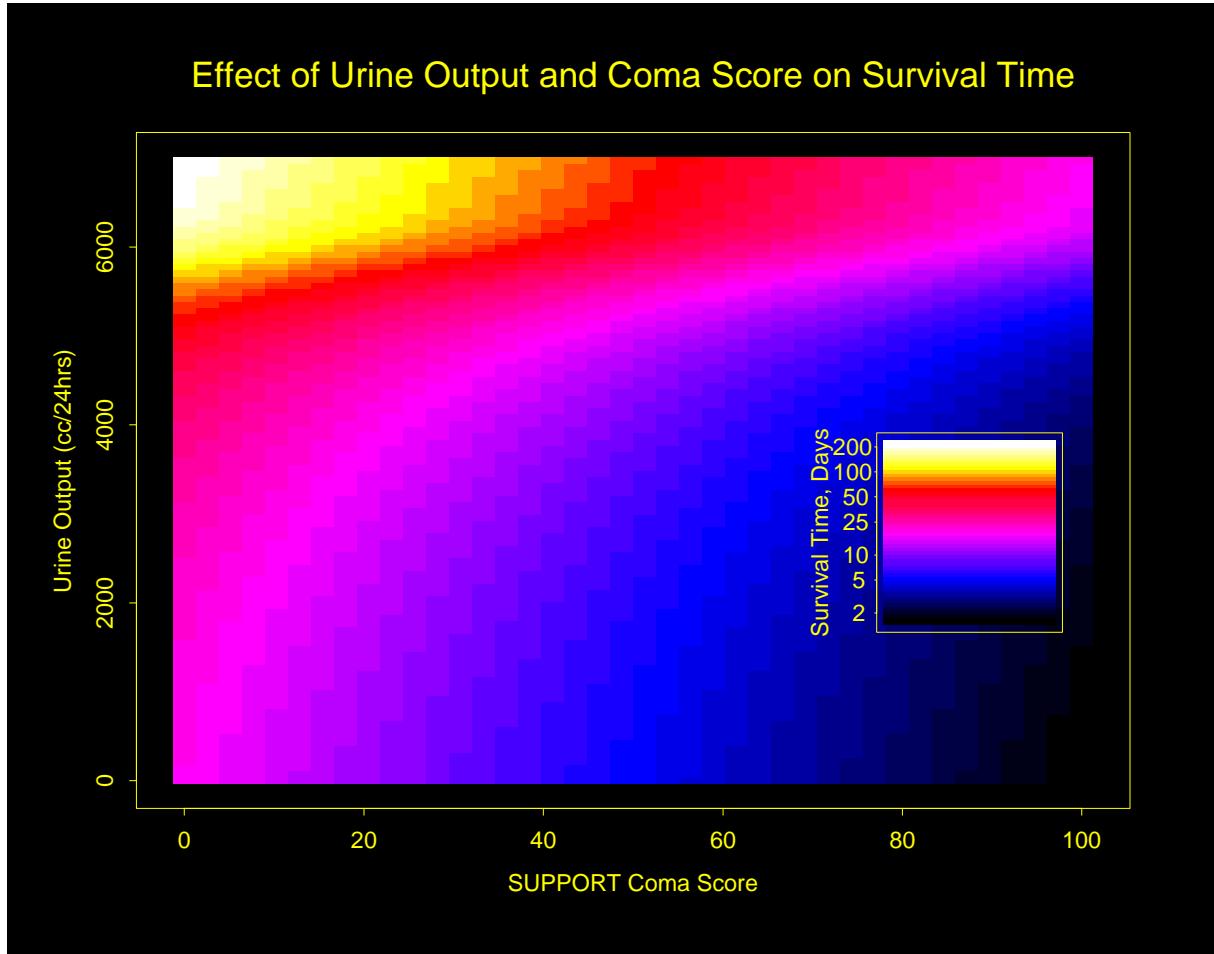


Figure 4.25: Estimated median survival time for critically ill adults

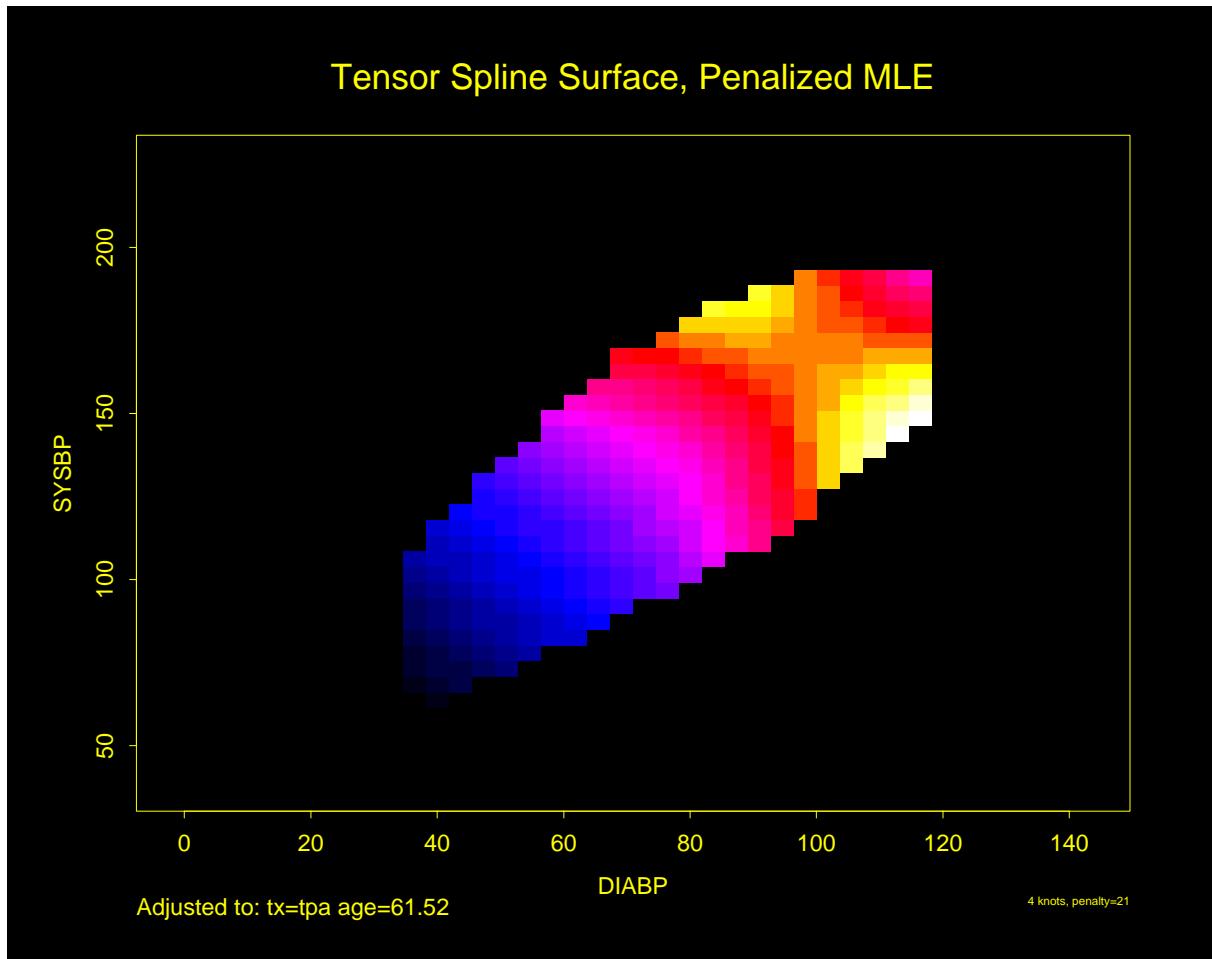


Figure 4.26: Logistic regression estimate of probability of a hemorrhagic stroke for patients in the GUSTO-I trial given t -PA, using a tensor spline of two restricted cubic splines and penalization (shrinkage). Dark (cold color) regions are low risk, and bright (hot) regions are higher risk.

that mean arterial blood pressure (roughly a 45° line) is what is most important over a broad range of blood pressures. At the far right one can see that pulse pressure (axis perpendicular to 45° line) may have an impact although a non-monotonic one.

4.4

Tables



What are tables for?

- Lookup of details
- Not for seeing trends
- For displaying a summary of a variable stratified by a truly categorical variable
- Not for summarizing how a variable changes across levels of a continuous independent variable

Since tables are for lookup, they can be complex. With modern media, a better way to think of a table is as a pop-up when viewing elements of a graph.

What to display in a table for different types of response variables:

- Binary variables: Show proportions first; they should be featured because they are normalized for sample size
 - Don't need to show both proportions (e.g., only show proportion of females)
 - Proportions are better than percents because of reduced confusion when speaking of percent difference (is it relative or absolute?) and because percentages such as 0.3% are often mistaken for 30% or 3%.
- Make logical choices for independent and dependent variables.
E.g., less useful to show proportion of males for patients who lived vs. those who died than to show proportion of deaths stratified by sex.
- Continuous response variables
 - to summarize distributions of raw data: 3 quartiles
recommended format: 35 **50** 67 or 35/**50**/67
 - summary statistics: mean or median and confidence limits (without assuming

normality of data if possible)

- Continuous independent (baseline) variables
 - Don't use tables because these requires arbitrary binning (categorization)
 - Use graphs to show smooth relationships
- Show number of missing values
- Add denominators when feasible
- Confidence intervals: in a comparative study, show confidence intervals for differences, not confidence intervals for individual group summaries

Table 4.1: Descriptive Statistics: Demographic and Clinical variables

	N		
Age	27	28	32 52
C-reactive protein	27	1.0	1.8 10.1
Fecal Calprotectin	26	128	754 2500
Gender	27		
Female		0.52	$\frac{14}{27}$
Location of colitis	27		
Left side		0.41	$\frac{11}{27}$
Middle		0.52	$\frac{14}{27}$
Right side		0.07	$\frac{2}{27}$

a b c represent the lower quartile a , the median b , and the upper quartile c for continuous variables.
 N is the number of non-missing values.

See also hbiostat.org/talks/rmedicine19.html#18.

Chapter 5

Statistical Inference

5.1

Overview

ABD6



- Inferential modes
 - hypothesis testing
 - relative model fit/relative support for hypotheses (likelihood ratios, Bayes factors)
 - estimation (including interval estimation; often more appropriate than hypothesis testing)
 - Bayesian probability of an effect in the right direction (more relevant to decision making; more actionable)
- Contrast inference with decision making:
 - acting *as if something is true* whether or not it is actually true
- Hypothesis testing is most useful for inferential “existence” questions
 - is indirect for other types of questions
- Majority of hypothesis tests are on a single point
- These place asymmetric importance on “special values” such as zero treatment

effect

- What does it mean to “not reject a hypothesis”?
 - often very little
- What does it mean to reject a hypothesis with a statistical test?
 - in statistics it means that *some* aspect of a model is under suspicion (e.g., normality assumption)
 - even if all other assumptions of the data model are satisfied, when the hypothesis involves a single point (e.g. zero effect), the alternative space is infinite so what have we learned about a *specific* alternative (e.g., that a treatment lowers blood pressure by 10 mmHg)?

Statistical hypothesis testing involves a model for data:

- Parametric tests have very specific models
- Nonparametric tests have semi-specific models without a distribution assumption
- Permutation tests make an assumption about the best data summarization (e.g., mean, and the mean may not be the best summary for a heavy-tailed data distribution)

This chapter covers parametric tests for the following reasons:

1. historical
2. they are very useful for sample size estimation
3. occasionally one has prior information that the raw data, or differences from pre to post, actually follow a normal distribution, and with large effects one can get quite significant results in very small samples with parametric tests
4. to show that Bayesian parametric tests make fewer assumptions about the data model

Nonparametric methods are covered in Chapter [7](#).

Back to data model—the primary basis for statistical inference:

- Assume a model for how the data were generated
- Model contains
 - main parameters of interest (e.g., means)
 - auxiliary parameters for other variables (e.g., confounders)
 - parameters for sources of between-subject variability
 - if parametric, a distribution function such as Gaussian (normal)
 - if nonparametric/semiparametric, a connection between distributions for different types of subjects (link function)
 - if longitudinal, a correlation pattern for multiple measurements within a subject
 - assumptions about censoring, truncation, detection limits if applicable
 - ...
- Example (2-sample t -test): $Y = \mu_0 + \delta[\text{treatment B}] + \epsilon$
 - μ_0 : unknown data-generating mean for treatment A
 - δ : unknown data-generating difference in means (B-A)
 - [treatment B]: an indicator variable (1 if observation is from treatment B, 0 if from treatment A)
 - ϵ : irreducible error, i.e., unaccountable subject-to-subject variation; biologic variability; has variance σ^2
 - **Primary goal:** uncover the hidden value of δ generating our dataset in the presence of noise ϵ
 - * higher $\sigma^2 \rightarrow$ larger $|\epsilon| \rightarrow$ harder to uncover δ (the signal) from the noise
 - * were ϵ always zero (no uncertainties), one *directly observes* δ and no statistical analysis is needed

- Rejecting a straw-man hypothesis implies that *some* aspect of the model is in doubt
 - that aspect may be the distribution or an assumption about equal variances, not the difference in means you care about

Common error: Using a two-stage approach to select which data model to use:

- Assumes the data contain rich enough information to lead to a correct decision
- Alters the operating characteristics of the final test
- Fails to realize that nonparametric tests have excellent power

Example: Testing normality to decide on whether to use a *t*-test vs. a Wilcoxon-Mann-Whitney two-sample rank test. A two-stage test with a test for normality in the first stage assumes that

1. the test for normality has power near 1.0 for our sample size
2. if the test rejects H_0 :normality, the magnitude of non-normality is worth bothering about
3. pre-testing for normality does not modify the type I error of the overall testing procedure
4. nonparametric tests are less efficient than parametric tests

In fact it may be that none of these assumptions is true (4. is very seldom true). As will be seen later, a full Bayesian model can be completely honest and provide exactly the right amount of caution:

- flexible parametric model allowing uncertainty about equal variances for groups A and B
- allows uncertainty about the normality assumption
- still results in inference about δ , properly more cautious (wider credible interval) because
 - we don't know if normality truly holds

- we don't know if equality of variance truly holds

5.1.1

Central Limit Theorem

- Assume observations are independent with the same distribution and have finite variance
- Assume that the sample size n can increase without bound
- Limiting distribution of the sample mean is normal

The CLT is frequently used to justify the use of parametric statistical tests and confidence intervals even when their assumptions are violated. **But** the CLT (and the t distribution) are much less helpful for computing confidence intervals and P -values than it seems¹¹³:

- since the standard deviation is unknown, one must estimate it while estimating the mean to use the CLT
- SD may be a completely inappropriate summary of dispersion (e.g., if one should have first log-transformed but computed SD on the raw scale)
- if the data distribution is asymmetric, the SD is not independent of the mean (so the t distribution does not hold) and the SD is not a good dispersion measure
- the sample size for which the CLT is an adequate approximation is unknown for any given situation
- example: log-normal distribution—the CLT is not accurate even for $n = 50000$ (see below)
- even when the CLT provides somewhat accurate P -values, it provides no comfort with regard to statistical power. Example: analyzing Y when you should have analyzed $\log(Y)$ will result in a devastating increase in type II error.

Example simulation to compute confidence-coverage when using the t -distribution con-

fidence interval for $n = 50,000$ when analyzing data from a log-normal distribution but not taking logs. We want the confidence limits to be such that the fraction of samples for which the true population mean is to the left of the lower limit is 0.025, and the fraction to the right of the upper limit to also be 0.025.

```

n      ← 50000
nsim ← 5000          # number of simulations
mul ← 0; sdl ← 1.65    # on log scale
mu ← exp(mul + sdl * sdl / 2)    # population mean on orig. scale
count ← c(lower=0, upper=0)
set.seed(1)
z ← qt(0.975, n - 1)    # t critical value (near 1.96)

for(i in 1 : nsim) {
  x ← exp(rnorm(n, mul, sdl))
  ci ← mean(x) + c(-1, 1) * z * sqrt(var(x) / n)
  count[1] ← count[1] + (ci[1] > mu)
  count[2] ← count[2] + (ci[2] < mu)
}
count / nsim    # non-coverage prob. in left and right tail

```

lower	upper
0.0182	0.0406

See stats.stackexchange.com/questions/204585 for more information and discussion.

5.2

Hypotheses

5.2.1

Scientific Hypotheses and Questions

Scientific questions are usually stated with a direction or with regard to an expected effect, not with respect to a null effect. Scientific questions often involve estimating a quantity of interest. Here are some examples:

- Does the risk of death increase when serum triglyceride increases?
- To what extent is mechanism x responsible for regulation of physiologic parameter y ?
- What is the average decrease in blood pressure when the dose of a drug goes from 0 to 10mg/day to 20mg/day?

5.2.2

Statistical Hypotheses

- Hypothesis: usually a statement to be judged of the form “population value = specified constant”
 - $\mu = 120\text{mmHg}$
 - $\mu_1 - \mu_2 = 0\text{mmHg}$
 - Correlation between wealth and religiosity = 0
- Null hypothesis is usually a hypothesis of no effect but can be $H_0 : \mu = \text{constant}$ or $H_0 : \text{Probability of heads} = \frac{1}{2}$;
 H_0 is often a straw man; something you hope to disprove
- Alternative hypothesis: H_1 ; e.g.: $H_1 : \mu \neq 120\text{mmHg}$

- One-sided hypothesis (tested by 1-tailed test): H_1 is an inequality in one direction ($H_1 : \mu > 120\text{mmHg}$)
- Two-sided hypothesis (2-tailed test, most common type): H_1 involves values away from the hypothesized value in either direction

5.3

Branches of Statistics

- Classical (frequentist or sampling statistics)
 - Emphasizes (overemphasizes?) hypothesis testing
 - Assumes H_0 is true and tries to amass evidence casting doubt on this assumption
 - Conceives of data as one of many datasets that *might* have happened; considers the process by which the sample arose
 - Inference is based on long-run operating characteristics not about direct evidence from the sample at hand
 - * probability of making an assertion of an effect if there is no effect
 - * proportion of the time that varying confidence intervals over replicates of the experiment cover the true unknown parameter; no statement about the chance that the current confidence interval covers the true parameter
 - See if data are consistent with H_0
 - Are data extreme or unlikely if H_0 is really true?
 - Proof by contradiction: if assuming H_0 is true leads to results that are “bizarre” or unlikely to have been observed, casts doubt on premise
 - Evidence summarized through a single statistic capturing a tendency of data, e.g., \bar{x}
 - Look at probability of getting a statistic as or more extreme than the calculated one (results as or more impressive than ours) if H_0 is true (the *P*-value)^a
 - If the statistic has a low probability of being more extreme we say that if H_0 is true we have acquired data that are very improbable, i.e., have witnessed a low probability event

^aWe could drop the “as” and just say “more extreme” because for continuous data the probability of getting a result exactly as extreme as ours is exactly zero.

- William Briggs in wmbriggs.com/public/Briggs.EverthingWrongWithPvalues.pdf described a basic flaw in the logic of P -value-guided hypothesis testing dating all the way back to Fisher:

A version of an argument given first by Fisher appears in every introductory statistics book. The original argument is this, [33]:

"Belief in a null hypothesis as an accurate representation of the population sampled is confronted by a logical disjunction: Either the null hypothesis is false, or the p-value has attained by chance an exceptionally low value."

A logical disjunction would be a proposition of the type "Either it is raining or it is not raining." Both parts of the proposition relate to the state of rain. The proposition "Either it is raining or the soup is cold" is a disjunction, but not a logical one because the first part relates to rain and the second to soup. Fisher's "logical disjunction" is evidently not a logical disjunction because the first part relates to the state of the null hypothesis and the second to the p-value. Fisher's argument can be made into a logical disjunction, however, by a simple fix. Restated: Either the null hypothesis is false and we see a small p-value, or the null hypothesis is true and we see a small p-value. Stated another way, "Either the null hypothesis is true or it is false, and we see a small p-value." The first clause of this proposition, "Either the null hypothesis is true or it is false", is a tautology, a necessary truth, which transforms the proposition to (loosely) "TRUE and we see a small p-value." Adding a logical tautology to a proposition does not change its truth value; it is like multiplying a simple algebraic equation by 1. So, in the end, Fisher's dictum boils down to: "We see a small p-value." In other words, in Fisher's argument a small p-value has no bearing on any hypothesis (any hypothesis unrelated to the p-value itself, of course). Making a decision about a parameter or data because the p-value takes any particular value is thus always fallacious: it is not justified by Fisher's argument, which is a non sequitur.

Ignoring all that and plunging ahead:

- P -value is a measure of *surprise* that is well described by Nicholas Maxwell^b: "A p value is a measure of how embarrassing the data are to the null hypothesis"
- Then evidence mounts against H_0 and we might reject it
- A failure to reject *does not* imply that we have gathered evidence in favor of H_0 — many reasons for studies to not be impressive, including small sample size (n)

^bData Matters: Conceptual Statistics for a Random World. Emeryville CA: Key College Publishing, 2004.

- Ignores *clinical* significance
- Is fraught with how to deal with multiplicity problems
 - * No principled recipe for how they should be handled
 - * Arise because
 - type I error is fixed at a number > 0
 - backward time ordering of information (transposed conditional)
 - * Evidence about one question is changed according to whether other questions are asked (regardless of their answers)
- Classical parametric statistics: assumes the data to arise from a certain distribution, often the normal (Gaussian distribution)
- Nonparametric statistics: does not assume a data distribution; generally looks at ranks rather than raw values
- Bayesian statistics:
 - Considers the sample data, not how it arose from a sequence of samples but rather the data generating mechanism for *this* sample
 - Computes the probability that a clinically interesting statement is true, e.g. that the new drug lowers population mean SBP by at least 5mmHg, given what we observed in the data
 - Instead of trying to amass evidence against a single hypothesized effect size, Bayes tries to uncover the hidden parameter generating the data aside from noise (e.g., treatment effect) whatever its value
 - * Provides evidence for *all possible values* of an unknown parameter
 - More natural and direct approach but requires more work
 - Because respects forward flow of time/information there is no need for nor availability of methods for correcting for multiplicity^c

^cBayesian inference assumes only that the prior distribution is “well calibrated” in the sense that one sticks to the pre-specified prior no matter what information is unveiled.

- Evidence about one question is not tilted by whether other questions are asked
 - Can formally incorporate knowledge from other studies as well as skepticism from a tough audience you are trying to convince to use a therapy
 - Starting to catch on (only been available for about 240 years) and more software becoming available
- Likelihood inference^d:
 - Considers the sample data, not how it arose
 - Akin to Bayesian but without the prior
 - Interval estimates are based on relative likelihood (from the likelihood function) and are called likelihood support intervals
 - For testing, allows both type I and type II errors $\rightarrow 0$ as $n \rightarrow \infty$, whereas with frequentist methods the type I error never shrinks as $n \rightarrow \infty$
 - This greatly reduces problems with multiplicities
 - Likelihood methods do not deal well with complex assertions (e.g., either the treatment reduces mortality by any amount or reduces blood pressure by at least 10 mmHg) and do not allow the use of external information
 - Bayesian and likelihood inference use the *likelihood principle*; frequentist inference does not
 - Likelihood principle: All of the evidence in a sample relevant to model parameters is in the likelihood function
 - If we want to know our current location, frequentist inference asks the following: If I am in Nashville, what fraction of routes to here involved the southern route I took? There are many paths to get where we are, and frequentists have to consider all possible relevant paths. Bayesian and likelihood inference states it differently: Where am I now? This involves an assessment of current evidence about my location. Asking “how did I get here?” (i.e., how did the data arise?) involves multiplicity issues that answering the simpler question does not.

^dA key thinker and researcher in the field is Richard Royall.

- Consider a sequentially monitored randomized experiment. Bayesians and likelihoodists can make infinitely many assessments of efficacy with no penalty. On the other hand, a frequentist must think the following way:

I am at the first interim analysis. I am going to make later assessments of efficacy so I need to discount the current assessment and be more conservative or I will spend all my α already.

...

I am at the fifth interim analysis. I made four previous efficacy assessments, and even though none of them mattered, I spent α so I need to discount the current assessment and be more conservative.

- We will deal with classical parametric and nonparametric statistical tests more than Bayesian methods just because of time and because of abundance of software for the former

5.4

Errors in Hypothesis Testing; P Values

- Can attempt to reject a formal hypothesis or just compute P -value
- Type I error: rejecting H_0 when it is true
 α is the probability of making this error (typically set at $\alpha = 0.05$ —for weak reasons)
 - To be specific: $\alpha = P(\text{asserting an effect exists when in fact it doesn't})$
 - So it is an assertion probability or false alarm probability like 1 minus specificity
 - It **is not** a false positive probability ($P(\text{effect}=0)$ given an assertion that it is nonzero) since α is based on an **assumption** that effect=0
- Type II error: failing to reject H_0 when it is false
 probability of this is β

		True state of H_0	
Decision		H_0 true	H_0 false
Reject H_0	Type I error (α)	Correct	
	Do Not Reject H_0	Correct	Type II error (β)

- Power: $1 - \beta$: probability of (correctly) rejecting H_0 when it is false

Within the frequentist framework of statistics there are two schools. One, the Neyman-Pearson school, believes that the type I error should be pre-set at α (say $\alpha = 0.05$) so that binary decisions can be made (reject/fail to reject). The other school due to Fisher believes that one should compute P -values and quote the result in the report or publication. This is the more popular approach, being less dichotomous.

- Simplest definition of α : probability that a P -value will be less than it

A P -value is something that can be computed without speaking of errors. It is the probability of observing a statistic as or more extreme than the observed one if H_0 is true, i.e., if the population from which the sample was randomly chosen had the

characteristics posited in the null hypothesis.^e The P -value to have the usual meaning assumes that the data model is correct.

5.4.1

Problems With Type I Error

- It's not really an "error"
 - Error = being wrong in asserting an effect
 - Type I = $P(\text{asserting effect } | H_0) = \text{"effect assertion probability"}$
- Frequentist designs attempt to preserve type I error
- This is not the probability of making a mistake in concluding an effect is present
- When $\alpha = 0.05$ the probability of asserting an effect when there is none never decreases even as $n \rightarrow \infty$
(statistical vs. clinical significance problem)
- α does not depend on any observed data. It is a pre-study concept.
- α increases because of chances you give data to be more extreme (multiplicity), not because of chances you give hypotheses to be false. Bayesian and likelihood approaches do not look at sampling (sample space giving rise to data extremes). See fharrell.com/post/bayes-seq
- Probability of making a mistake in asserting an effect, given the data, is one minus the Bayesian posterior probability of efficacy

5.4.2

Misinterpretation of P -values

P -values have come under extreme criticism since 2000, partially because they are often misinterpreted. Greenland *et al.*³⁸ is the best paper that summarizes the misinterpreta-

^eNote that Rosner's Equation 7.4 in his section 7.3 is highly problematic. Classifications of "significant" or "highly significant" are arbitrary, and treating a P -value between 0.05 and 0.1 as indicating a "trend towards significance" is bogus. If the P -value is 0.08, for example, the 0.95 confidence interval for the effect includes a "trend" in the opposite (harmful) direction.

tions and explains them. Some quotes from this paper are below, with their explanations for the first two.

1. **The P value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave $P = 0.01$, the null hypothesis has only a 1% chance of being true; if instead it gave $P = 0.40$, the null hypothesis has a 40% chance of being true.** No! The P value assumes the test hypothesis is true—it is not a hypothesis probability and may be far from any reasonable probability for the test hypothesis. The P value simply indicates the degree to which the data conform to the pattern predicted by the test hypothesis and all the other assumptions used in the test (the underlying statistical model). Thus $P = 0.01$ would indicate that the data are not very close to what the statistical model (including the test hypothesis) predicted they should be, while $P = 0.40$ would indicate that the data are much closer to the model prediction, allowing for chance variation.
2. **The P value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.** No! This is a common variation of the first fallacy and it is just as false. To say that chance alone produced the observed association is logically equivalent to asserting that every assumption used to compute the P value is correct, including the null hypothesis. Thus to claim that the null P value is the probability that chance alone produced the observed association is completely backwards: The P value is a probability computed assuming chance was operating alone. The absurdity of the common backwards interpretation might be appreciated by pondering how the P value, which is a probability deduced from a set of assumptions (the statistical model), can possibly refer to the probability of those assumptions. Note: One often sees “alone” dropped from this description (becoming “the P value for the null hypothesis is the probability that chance produced the observed association”), so that the statement is more ambiguous, but just as wrong.
3. **A significant test results ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.** No!
4. **A nonsignificant test results ($P > 0.05$) means that the test hypothesis is true or should be accepted.** No!
5. **A large P value is evidence in favor of the test hypothesis.** No!

6. A null-hypothesis P value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated. No!
7. Statistical significance indicates a scientifically or substantively important relation has been detected. No!
8. Lack of statistical significance indicates that the effect size is small. No!
9. The P value is the chance of our data occurring if the test hypothesis is true; for example, $P = 0.05$ means that the observed association would occur only 5% of the time under the test hypothesis. No!
10. If you reject the test hypothesis because $P \leq 0.05$, the chance you are in error (the chance your “significant finding” is a false positive) is 5%. No!
11. $P = 0.05$ and $P \leq 0.05$ mean the same thing. No!
12. P values are properly reported as inequalities (e.g., report “ $P < 0.02$ ” when $P = 0.015$ or report $P > 0.05$ when $P = 0.06$ or $P = 0.70$). No!
13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance. No!
14. One should always use two-sided P values. No!
15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all $P > 0.05$), the overall evidence supports the hypothesis. No!
16. When the same hypothesis is tested in two different populations and the resulting P values are on opposite sides of 0.05, the results are conflicting. No!
17. When the same hypothesis is tested in two different populations and the same P values are obtained, the results are in agreement. No!
18. If one observes a small P value, there is a good chance that the next study will produce a P value at least as small for the same hypothesis. No!
19. The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size. No!

20. An effect size outside the 95% confidence interval has been refuted (or excluded) by the data. No!
21. If two confidence intervals overlap, the difference between the two estimates or studies is not significant. No!
22. An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval. No!
23. If one 95% confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one. No!
24. If you accept the null hypothesis because the null P value exceeds 0.05 and the power of your test is 90%, the chance you are in error (the chance that your finding is a false negative) is 10%. No!
25. If the null P value exceeds 0.05 and the power of this test is 90% at an alternative, the results support the null over the alternative. . . counterexamples are easy to construct . . .

5.5

Interval Estimation

5.5.1

Frequentist Confidence Intervals

A $1 - \alpha$ two-sided confidence interval is an interval computed such that

- if the experiment were repeated N times one would expect $N \times (1 - \alpha)$ of the recomputed varying intervals to contain the true unknown quantity of interest
- equivalently the set of all unknown population parameter values that if null hypothesized one would not reject that null hypothesis at the α level in a two-sided test
 - e.g. when estimating the population mean μ , the set μ_0 such that the test $H_0 : \mu = \mu_0$ has a P -value $> \alpha$

For this reason confidence intervals may better be called *compatibility intervals*.

Pros:

- The P -value can be computed from the confidence interval but not vice-versa, so confidence limits have more information
- Confidence intervals do not allow the “absence of evidence is not evidence of absence error”
 - large P -values can come from small n or large σ^2
 - these make confidence intervals wide, giving a rational sense of uncertainty
 - A confidence interval that is compatible with both large benefit and large detriment indicates that we don’t know much
 - * large P -value means nothing more than “get more data”

Cons:

- Confidence intervals have only a long-run interpretation over many experiments
- They don't provide a probability statement about whether a single interval includes the true population value
 - In the frequentist world, the probability that the parameter is in a given interval is either zero or one
- They are often taken to provide a measure of precision of a statistical estimate, but they're not really that either
- The experimenter controls the long-run inclusion probability $1 - \alpha$ and gets interval endpoints, but is often more interested in the probability that the population effect is in a pre-chosen fixed interval
- It is very difficult to make confidence intervals incorporate known uncertainties (e.g., amount of non-normality), making them often too narrow (overoptimistic)

5.5.2

Bayesian Credible Intervals

- Credible intervals have the interpretation that most researchers seek when they compute confidence intervals
- A $1 - \alpha$ credible interval is an interval $[a, b]$ (computed, under a certain prior distribution encapsulating prior knowledge about the parameter μ) so that
$$P(a \leq \mu \leq b | \text{data}) = 1 - \alpha$$

Pros:

- Pertains to the single, current dataset and does not provide just long-run operating characteristics
- Provides a true probability statement even when the experiment could never be repeated

- Is symmetric with Bayesian posterior probabilities of pre-chosen limits
 - Researchers can specify a, b and then compute whatever probability it is that the unknown parameter is in that interval
- The credible interval can take into account all sorts of uncertainties (e.g., non-normality) that make it (correctly) wider

Cons:

- One must specify a prior distribution

5.6

One Sample Test for Mean



5.6.1

Frequentist Method

- Assuming continuous response from a normal distribution
- One sample tests for $\mu = \text{constant}$ are unusual except when data are paired, e.g., each patient has a pre- and post-treatment measurement and we are only interested in the mean of post - pre values
- t tests in general:

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of numerator}}$$

- The standard deviation of a summary statistic is called its *standard error*, which is the \sqrt of the variance of the estimate
- The one-sample t statistic for testing a single population mean against a constant μ_0 ($H_0: \mu = \mu_0$; often $\mu_0 = 0$) is

$$t = \frac{\bar{x} - \mu_0}{se}$$

where $se = \frac{s}{\sqrt{n}}$, is the standard error of the mean (SEM) and \bar{x} is the sample mean

- When your data comes from a normal distribution and H_0 holds, the t ratio **statistic** follows the **t distribution**
- With small sample size (n), the t ratio is unstable because the sample standard deviation (s) is not precise enough in estimating the population standard deviation (σ ; we are assuming that σ is unknown)
- This causes the t distribution to have heavy tails for small n
- As $n \uparrow$ the t distribution becomes the normal distribution with mean zero and standard deviation one

- The parameter that defines the particular t distribution to use as a function of n is called the *degrees of freedom* or d.f.
- d.f. = $n - \text{number of means being estimated}$
- For one-sample problem d.f. = $n - 1$
- Symbol for distribution t_{n-1}

```
x <- seq(-3, 3, length=200)      # Fig. 5.1
w <- list(Normal      = list(x=x, y=dnorm(x)),
          't (50 df)' = list(x=x, y=dt(x, 50)),
          't (5 df)'  = list(x=x, y=dt(x, 5)),
          't (2 df)'  = list(x=x, y=dt(x,2)))
labcurve(w, pl=TRUE, keys='lines', col=c(1,2,4,5), lty=c(2,1,1,1),
         xlab=expression(x), ylab='')
```

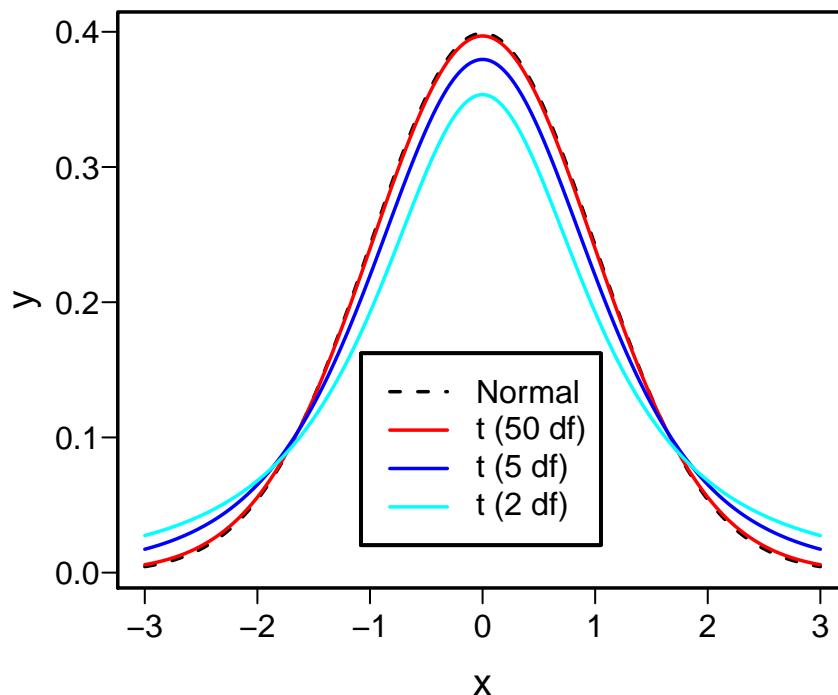


Figure 5.1: Comparison of probability densities for t_2 , t_5 , t_{50} , and normal distributions

- Two-tailed P -value: probability of getting a value from the t_{n-1} distribution as big or bigger in absolute value than the absolute value of the observed t ratio
- Computer programs can compute the P -value given t and n .^f R can compute all probabilities or critical values of interest. See the help files for `pt`, `pnorm`, `pf`, `pchisq`.

^fR has the function `pt` for the cumulative distribution function for the t distribution, so the 2-tailed P -value would be obtained using `2*(1-pt(abs(t),n-1))`.

- don't say " $P < \text{something}$ " but instead $P = \text{something}$
- In the old days tables were used to provide *critical values* of t , i.e., a value c of t such that $\text{Prob}[|t| > c] = \alpha$ for "nice" α such as 0.05, 0.01.
- Denote the critical value by $t_{n-1;1-\alpha/2}$ for a 2-tailed setup
- For large n (say $n \geq 500$) and $\alpha = 0.05$, this value is almost identical to the value from the normal distribution, 1.96
- Example: We want to test if the mean tumor volume is 190 mm^3 in a population with melanoma, $H_0 : \mu = 190$ versus $H_1 : \mu \neq 190$.

$$\bar{x} = 181.52, s = 40, n = 100, \mu_0 = 190$$

$$t = \frac{181.52 - 190}{40/\sqrt{100}} = -2.12$$

$t_{99,975} = 1.984 \rightarrow \text{reject at arbitrary } \alpha = .05 \text{ if using Neyman - Pearson paradigm}$

$$P = 0.037$$

```
xbar  ← 181.52
s      ← 40
n      ← 100
mu0   ← 190
tstat ← (xbar - mu0) / (s / sqrt(n))
pval  ← 2 * (1 - pt(abs(tstat), n - 1))
c(tstat=tstat, pval=pval)
```

<pre>tstat pval -2.12000000 0.03650607</pre>
--

5.6.2

Bayesian Methods

- All aspects of Bayesian probabilistic inference follow from the general form of Bayes' rule allowing for the response Y to be continuous:
The probability density function for the unknown parameters given the data and prior is proportional to the density function for the data multiplied by the density function for the prior^g

^gWhen Y is a discrete categorical variable we use regular probabilities instead of densities in Bayes' formula. The density at x is the limit as $\epsilon \rightarrow 0$ of the probability of the variable being in the interval $[x, x + \epsilon]$ divided by ϵ .

- The Bayesian counterpart to the frequentist t -test-based approach can use the same model
- But brings extra information for the unknowns— μ, σ —in the form of prior distributions for them
- Since the raw data have a somewhat arbitrary scale, one frequently uses nearly flat “weakly-informative” priors for these parameters
 - But it would be easy to substitute a prior for μ that rules out known-to-be impossible values or that assumes very small or very large μ are very unlikely
- **Note:** the assumption of a Gaussian distribution for the raw data Y is a strong one
 - Bayes makes it easy to relax the normality assumption but still state the inference in familiar terms (e.g., about a population mean)^h
 - See later for an example
- For now assume that normality is known to hold, and use relatively uninformative priors

Bayesian Software

In most Bayesian analyses that follow we use the R `brms` package by Paul-Christian Bürkner that is based on the general `Stan` system because `brms` is easy to use, makes good default choices for priors, and uses the same notation as used in frequentist models in Rⁱ.

Except for the one-sample proportion example in the next section, our Bayesian calculations are general and do not assume that the posterior distribution has an analytic solution. Statistical samplers (Markov chain Monte Carlo, Gibbs, and many variants) can sample from the posterior distribution only by knowing the part of Bayes' formula that is the simple product of the data likelihood and the prior, without having to integrate to get a marginal distribution^j. We are generating 4000 samples from the posterior distribution (4000 random draws) in this chapter. When high precision of

^hSome analysts switch to trimmed means in the presence of outliers but it is hard to interpret what they are estimating.

ⁱThanks to Nathan James of the Vanderbilt Department of Biostatistics for providing `brms` code for examples in this chapter.

^jThe simple product is proportional to the correct posterior distribution and just lacks a normalizing constant that makes the posterior probabilities integrate to 1.0.

posterior probabilities is required, one can ensure that all probability calculations have a margin of simulation error of < 0.01 by using 10,000 samples, or margin of error < 0.005 by taking 40,000 draws^k.

Example

To introduce the Bayesian treatment for the one-sample continuous Y problem consider the following made-up dataset that has an “outlier”:

```
98 105 99 106 102 97 103 132
```

- Start with regular frequentist analysis
- Most interested in confidence interval
- Also test $H_0 : \mu = 110$

Five measures of dispersion are computed^l.

```
y ← c(98, 105, 99, 106, 102, 97, 103, 132)
median(y)
```

```
[1] 102.5
```

```
sd(y)
```

```
[1] 11.28526
```

```
mean(abs(y - mean(y)))
```

```
[1] 6.875
```

```
mean(abs(y - median(y)))
```

```
[1] 6.25
```

```
GiniMd(y) # Gini's mean difference is more robust to outliers
```

```
[1] 10.85714
```

```
# It is the mean absolute difference over all pairs of observations
median(abs(y - median(y))) # Even more robust, not as efficient
```

^kThese numbers depend on the sampler having converged and the samples being independent enough so that the number of draws is the effective sample size. Stan and brms have diagnostics that reveal the effective sample size in the face of imperfect posterior sampling.

^lIn a standard normal(0,1) distribution, the measures are, respectively, 1.0, 0.8, 0.8, 1.13, and 0.67.

```
[1] 3.5
```

```
t.test(y, mu=110)
```

```
One Sample t-test

data: y
t = -1.1905, df = 7, p-value = 0.2727
alternative hypothesis: true mean is not equal to 110
95 percent confidence interval:
 95.81528 114.68472
sample estimates:
mean of x
 105.25
```

Now consider a Bayesian counterpart. A one-sample problem is a linear model containing only an intercept, and the intercept represents the overall unknown data-generating (population) mean of Y . In R an intercept-only model has ~ 1 on the right-hand side of the model formula.

For the prior distribution for the mean we assume a most likely value (mean of distribution, since symmetric) of 150 and a standard deviation of 50, indicating the unknown mean is likely to be between 50 and 250. This is a weakly informative prior.

```
# Tell brms/Stan to use all available CPU cores
options(mc.cores=parallel::detectCores())

require(brms)

d <- data.frame(y)
priormu <- prior(normal(150,50), class='Intercept')
f <- brm(y ~ 1, family=gaussian, prior=priormu, data=d, seed=1)
```

See which prior distributions are being assumed. The data model is Gaussian. Also display parameter distribution summaries. Estimate represents the mean of the posterior distribution.

```
prior_summary(f)
```

	prior	class	coef	group	resp	dpar	nlpar	bound
1	normal(150, 50)	Intercept						
2	student_t(3, 0, 10)		sigma					

```
f
```

```
Family: gaussian
Links: mu = identity; sigma = identity
Formula: y ~ 1
```

```

Data: d (Number of observations: 8)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept   105.53      4.46    96.50   114.35  1.00     1893     1531

Family Specific Parameters:
    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma      12.18      3.38     7.56   20.22  1.00     1807     1941

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

```

```

draws <- as.data.frame(f)
mu     <- draws$b_Intercept
sigma <- draws$sigma
length(mu)

```

```
[1] 4000
```

- Note how Bayes provides uncertainty/precision information about σ
- Credible intervals are quantiles of posterior samples, e.g. we can duplicate the above credible interval (CI) for μ (the intercept) using the following:

```
quantile(mu, c(.025, 0.975))
```

```

 2.5%    97.5%
96.50034 114.35211

```

- Compare 0.95 credible interval with the 0.95 frequentist confidence interval above
- Compare posterior means for μ, σ with the point estimates from the traditional analysis
 - Posterior modes (the most likely values) may be more relevant^m. These are printed below along with sample estimates. The posterior mode is computed by fitting a nonparametric kernel density estimator to the posterior draws for the parameter of interest and finding the peak. The number of draws needed to compute the mode accurately is more than the number we are using here.

^mThe sample mean is the maximum likelihood estimate (MLE) and the sample standard deviation is, except for a factor of $\frac{n-1}{n}$, the maximum likelihood estimate. When priors are flat, MLEs equal posterior modes

```
# Function to compute posterior mode given draws
pmode <- function(x) {
  z <- density(x)
  z$x[which.max(z$y)]
}
n <- length(y)
c(mean(y), sd(y), sd(y)*sqrt((n-1)/n))
```

```
[1] 105.25000 11.28526 10.55640
```

```
c(pmode(mu), pmode(sigma))
```

```
[1] 105.85115 10.98257
```

- Instead of a hypothesis test we compute direct evidence for μ exceeding 110
 - posterior probability that $\mu > 110$ given the data and priors is approximated (to within simulation error) by the proportion of posterior draws for μ for which the value of μ exceeded 110
 - define a “probability operator” P that is just the proportion

```
P <- mean
P(mu > 110) # compare to 1-tailed p-value: 0.136
```

Here are posterior distributions for the two parameters along with convergence diagnostics

```
plot(f)
```

- Normality is a strong assumption
- Heavy tails can hurt validity of the estimate of the mean, uncertainty intervals, and P -values
- Easy to allow for heavy tails by adding a single parameter to the modelⁿ
- Assume the data come from a t distribution with unknown degrees of freedom ν
 - Use of t for the raw data distribution should not be confused with the use of the same t distribution for computing frequentist probabilities about the sample mean.

ⁿOne could use a data distribution with an additional parameter that allows for asymmetry of the distribution. This is not addressed here.

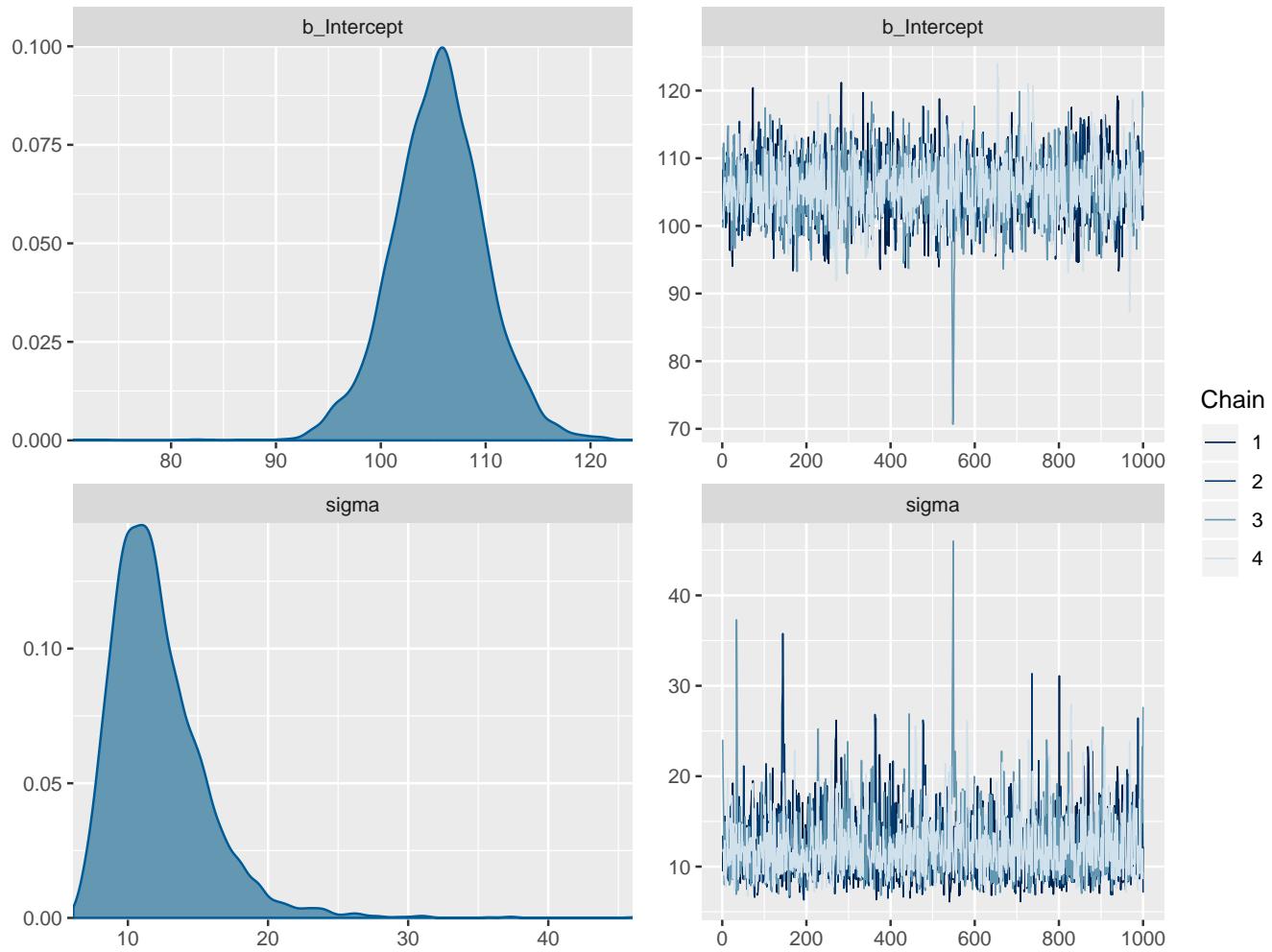


Figure 5.2: Posterior distributions for μ and σ using a normal data model with weak priors (left panels), and convergence diagnostics for posterior sampling (right panels)

- John Kruschke has championed this Bayesian t -test approach in [BEST](#) (Bayesian Estimation Supersedes the t -Test)
- $\nu > 20 \rightarrow$ almost Gaussian
- Have a prior for ν that is a gamma distribution with parameters $\alpha = 2, \beta = 0.1$ with ν constrained to be > 1
- Prior $P(\nu > 20)$ is (in R code) `pgamma(20, 2, 0.1, lower.tail=FALSE)` which is 0.41. So our prior probability that normality approximately holds is a little less than $\frac{1}{2}$.

```
g ← brm(y ~ 1, family=student, prior=priormu, data=d, seed=2)
```

```
prior_summary(g)
```

	prior	class	coef	group	resp	dpar	nlpar	bound
1	normal(150, 50)	Intercept						
2	gamma(2, 0.1)		nu					
3	student_t(3, 0, 10)			sigma				

```
g
```

```

Family: student
Links: mu = identity; sigma = identity; nu = identity
Formula: y ~ 1
Data: d (Number of observations: 8)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept 103.75     3.72    97.00   111.99 1.00     1722     1619
```

```

Family Specific Parameters:
Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     9.05      3.79     3.21    17.53 1.00     1400     1990
nu        14.30     12.87    1.47    47.98 1.00     1502     1585
```

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

```
draws ← as.data.frame(g)
mu     ← draws$b_Intercept
P(mu > 110)
```

```
[1] 0.054
```

```
plot(g)
nu     ← draws$nu
snu    ← c(mean(nu), median(nu), pmode(nu))
ssd    ← c(sd(y), pmode(sigma), pmode(draws$sigma))
fm     ← function(x) paste(round(x, 2), collapse=', ')
```

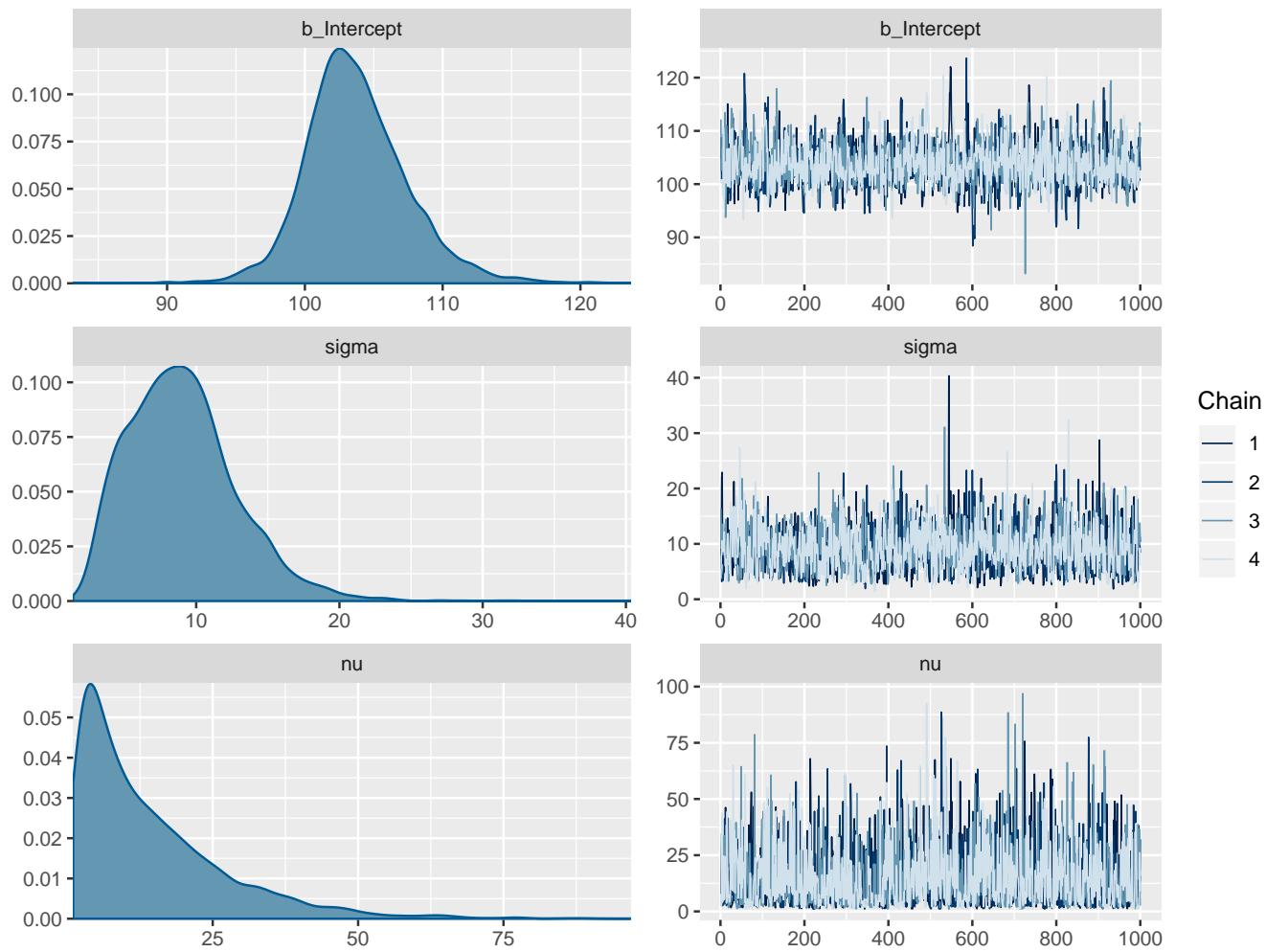


Figure 5.3: Posterior distributions for μ, σ, ν for a data model that is the t distribution with ν d.f. (left panels), and convergence diagnostics (right panels)

- Posterior means discount the high outlier a bit and shows lower chance of $\mu > 110$
- The credible interval for μ is significantly smaller than the confidence interval due to allowance for heavy tails
- The posterior mean, median, and mode of ν (14.3, 10.37, 3.97) provide evidence that the data come from a distribution that is heavier tailed than the normal. Posterior $P(\nu > 20) = 0.25$ which is essentially the probability of approximate normality under a t data model.
- The traditional SD estimate, posterior median SD assuming normal Y , and posterior median SD assuming Y has a t -distribution are respectively 11.29, 10.98, 8.8. The ordinary SD is giving too much weight to the high outlier.

This latter analysis properly penalizes for not knowing normality in Y to hold (i.e., not knowing the true value of ν).

Decoding the Effect of the Prior

- One can compute the effective number of observations added or subtracted by using, respectively, an optimistic or a skeptical prior
- This is particularly easy when the prior is normal, the data model is normal, and the data variance is known

- Let

μ_0 = prior mean

σ_0 = prior standard deviation

\bar{Y} = sample mean of the response variable

σ = population SD of Y

n = sample size for \bar{Y}

- Then the posterior variance of μ is

$$\sigma_p^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2 n}}$$

- The posterior mean is

$$\mu_p = \frac{\sigma_p^2}{\sigma_0^2} \times \mu_0 + \frac{\sigma_p^2}{\frac{\sigma_0^2}{n}} \times \bar{Y}$$

- For a given \bar{Y} , sample size n and posterior $P(\mu > 110)$, what sample size m would yield the same posterior probability under a flat prior?
- For a flat prior $\sigma_0 = \infty$ so the posterior variance of μ would be $\frac{\sigma^2}{n}$ and the posterior mean would be \bar{Y}
- Equating $P(\mu > 110)$ for the weakly informative prior at sample size n to that from noninformative prior at a sample size m is the same as setting $P(\mu < 110)$ to be equal
- The probability that a Gaussian random variable with mean a and standard deviation b is less than 110 is $\Phi(\frac{110-a}{b})$ where Φ is the standard normal cumulative distribution function
- Solve for m for a variety of n, \bar{Y}, σ

```

calc <- function(n, m, ybar, sigma, mu0, sigma0, cutoff) {
  vpost1 <- 1 / ((1 / (sigma0^2)) + 1 / ((sigma^2) / n))
  mupost1 <- mu0 * vpost1 / (sigma0 ^ 2) + ybar * vpost1 / ((sigma ^ 2) / n)
  vpost2 <- (sigma ^ 2) / m
  (110 - mupost1) / sqrt(vpost1) - (110 - ybar) / sqrt(vpost2)
}

# For a given n, ybar, sigma solve for m to get equal post. prob.
m <- function(n, ybar, sigma, sigma0=50, cutoff=110)
  round(uniroot(calc, interval=c(n / 2, 2 * n), n=n, ybar=ybar, sigma=sigma,
    mu0=150, sigma0=sigma0, cutoff=cutoff)$root, 1)
# Make sure that m=n when prior variance is huge
m(8, 100, 10, sigma0=50000)

```

[1] 8

```

# From here on use the original prior standard deviation of 50
# Now compute m when n=8 using our original prior assuming sigma=10 ybar=105
m(8, 105, 10)

```

[1] 7.3

```

# What about for two other sigmas
m(8, 105, 5)

```

[1] 7.8

```

m(8, 105, 15)

```

[1] 6.6

```
# What about two other ybar
m(8, 80, 10)
```

```
[1] 7.9
```

```
m(8, 120, 10)
```

```
[1] 8.3
```

```
# What if n were larger
m(15, 105, 10)
```

```
[1] 14.3
```

```
m(30, 105, 10)
```

```
[1] 29.3
```

```
m(300, 105, 10)
```

```
[1] 299.3
```

```
m(3000, 105, 10)
```

```
[1] 2999.3
```

- Typical effect of prior in this setting is like adding or dropping one observation
- Simpler example where \bar{Y} is irrelevant: $\sigma = 1, \mu_0 = 0$ and posterior probability of interest is $P(\mu > 0)$
- Vary the prior variance σ_0^2 and for each prior variance compute the prior probability that $\mu > 1$ (with prior centered at zero, lower variance \rightarrow lower $P(\mu > 1)$)

```
z ← list()
n ← seq(1, 100, by=2)
for(v in c(.05, .1, .25, .5, 1, 4, 100))
  z[[paste0('v=', v, ', P(mu>1)=',
            format(1 - pnorm(sqrt(1 / v)), digits=3, scientific=1))]] ←
    list(x=n, y=0.5 * (n + sqrt(n^2 + 4 * n / v)) - n)

labcurve(z, pl=TRUE, xlab='Sample Size With No Skepticism',
          ylab='Extra Subjects Needed Due to Skepticism', adj=1)
```

- For typical priors the effect of skepticism is like dropping 3 observations, and this is not noticeable when $n > 20^{\circ}$
- **Note:** For a specific problem one can just run the `brm` function again with a non-informative prior to help judge the effect of the informative prior

^oSee [here](#) for the derivation.

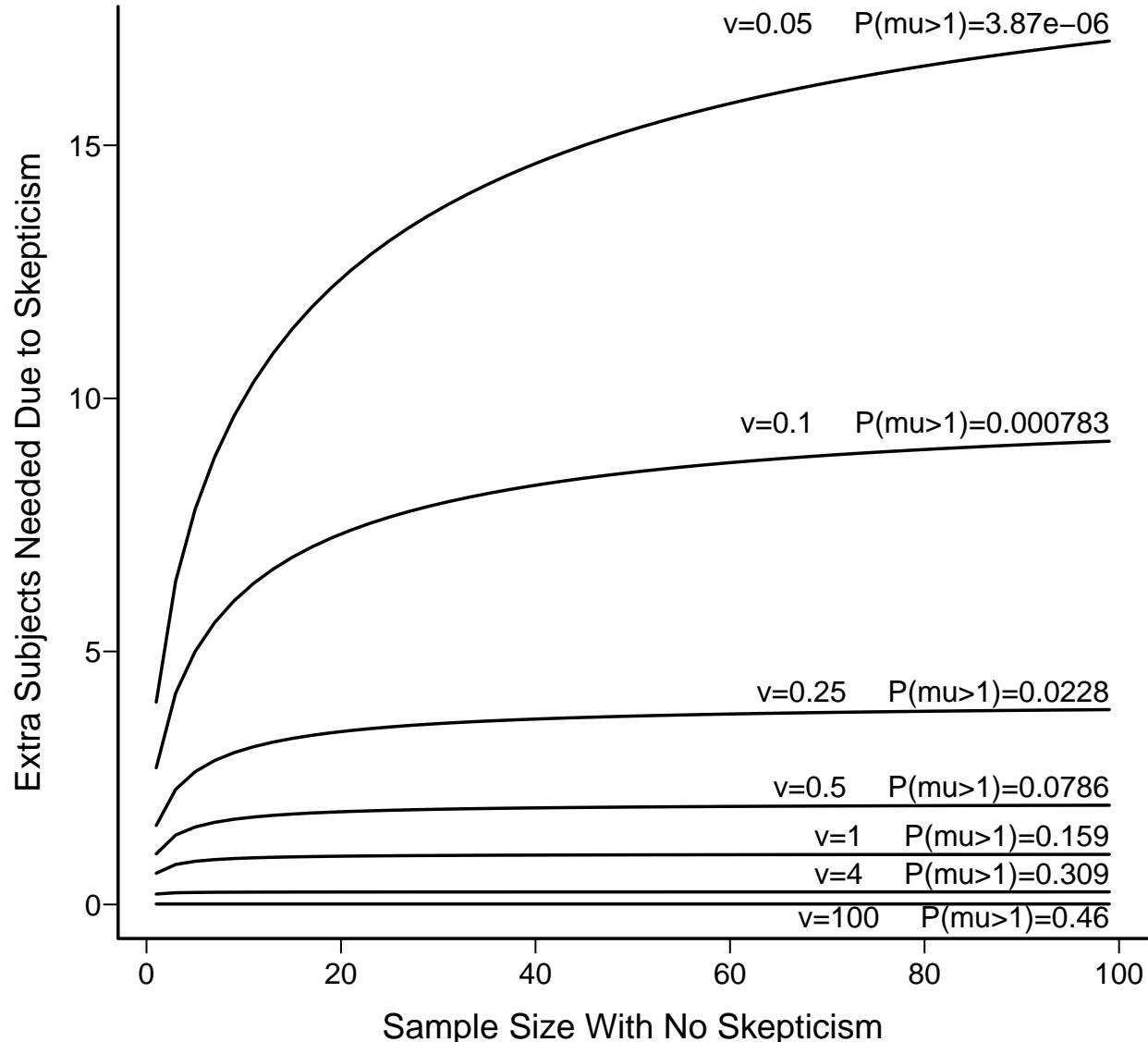


Figure 5.4: Effect of discounting by a skeptical prior with mean zero and variance v : the increase needed in the sample size in order to achieve the same posterior probability of $\mu > 0$ as with the flat (non-informative) prior. Prior variance $v=0.05$ corresponds to a very skeptical prior, given almost no chance to a large μ ($\mu > 1$).

5.6.3

Power and Sample Size

- Bayesian power is the probability of hitting a specified large posterior probability and is usually obtained by simulation
- Frequentist power, though often arbitrary (and inviting optimistic single point values for an effect to detect) is easier to compute and will provide a decent approximation for Bayesian sample size calculations when the main prior is weakly informative
- Frequentist power \uparrow when
 - allow larger type I error (α ; trade-off between type I and II errors)
 - true μ is far from μ_0
 - $\sigma \downarrow$
 - $n \uparrow$
- Power for 2-tailed test is a function of μ, μ_0 and σ only through $|\mu - \mu_0|/\sigma$
- Sample size to achieve $\alpha = 0.05$, power = 0.9 is approximately

$$n = 10.51 \left[\frac{\sigma}{\mu - \mu_0} \right]^2$$

- Some power calculators are at statpages.info/#Power
- PS program by Dupont and Plummer <http://biostat.mc.vanderbilt.edu/PowerSampleSize>
- Example: The mean forced expiratory volume (FEV) in a population of asthmatics is 2.5 liters per second and the population standard deviation is assumed to be 1. Determine the number of subjects needed if a new drug is expected to increase FEV to 3.0 liters per second ($\alpha = .05, \beta = 0.1$)

$$\mu = 2.5, \mu_0 = 3, \sigma = 1$$

$$n = 10.51 \left[\frac{1}{3 - 2.5} \right]^2 = 42.04$$

- Rounding up, we need 43 subjects to have 0.9 power (42 subjects would have less than 0.9 power)

```

sigma <- 1
mu     <- 2.5
mu0    <- 3
n      <- 10.51 * (1 / (mu - mu0)) ^ 2
# General formula for approximate power of 1-sample t-test
# Approximate because it uses the normal distribution throughout,
# not the t distribution
alpha <- 0.05
power <- 0.9
delta <- mu - mu0
za    <- qnorm(1 - alpha / 2)
zb    <- qnorm(power)
n     <- ((za + zb) * sigma / delta) ^ 2
c(alpha=alpha, power=power, delta=delta, za=za, zb=zb, n=n)

```

alpha	power	delta	za	zb	n
0.050000	0.900000	-0.500000	1.959964	1.281552	42.029692

A slightly more accurate estimate can be obtained using the t distribution, requiring iterative calculations programmed in R packages such as pwr.

```

# Make sure pwr package is installed
require(pwr)

```

```
pwr.t.test(d = delta / sigma, power = 0.9, sig.level = 0.05, type='one.sample')
```

```
One-sample t test power calculation
```

```

n = 43.99548
d = 0.5
sig.level = 0.05
power = 0.9
alternative = two.sided

```

5.6.4

Confidence Interval

A 2-sided $1 - \alpha$ confidence interval for μ under normality for Y is

$$\bar{x} \pm t_{n-1,1-\alpha/2} \times se$$

The t constant is the $1 - \alpha/2$ level critical value from the t -distribution with $n - 1$ degrees of freedom. For large n it equals 1.96 when $\alpha = 0.05$.

An incorrect but common way to interpret this is that we are 0.95 confident that the unknown μ lies in the above interval. The exact way to say it is that if we were able to repeat the same experiment 1000 times and compute a fresh confidence interval for μ from each sample, we expect 950 of the samples to actually contain μ . The confidence level is about the procedure used to derive the interval, not about any one interval. Difficulties in providing exact but still useful interpretations of confidence intervals has driven many people to Bayesian statistics.

The 2-sided $1 - \alpha$ CL includes μ_0 if and only if a test of $H_0 : \mu = \mu_0$ is not rejected at the α level in a 2-tailed test.

- If a 0.95 CL does not contain zero, we can reject $H_0 : \mu = 0$ at the $\alpha = 0.05$ significance level

$1 - \alpha$ is called the *confidence level* or *confidence coefficient*, but it is better to refer to *compatibility*

5.6.5

Sample Size for a Given Precision

ABD14.7

There are many reasons for preferring to run estimation studies instead of hypothesis testing studies. A null hypothesis may be irrelevant, and when there is adequate precision one can learn from a study regardless of the magnitude of a *P*-value. A nearly universal property of precision estimates is that, all other things being equal, increasing the sample size by a factor of four improves the precision by a factor of two.

- May want to estimate μ to within a margin of error of $\pm\delta$ with 0.95 confidence^p
- “0.95 confident” that a confidence interval includes the true value of μ
- If σ were known but we still used the *t* distribution in the formula for the interval, the confidence interval would be $\bar{x} \pm \delta$ where

$$\delta = \frac{t_{n-1, 1-\alpha/2}\sigma}{\sqrt{n}}$$

^pAdcock¹ presents both frequentist and Bayesian methods and for precision emphasizes solving for n such that the probability of being within ϵ of the true value is controlled, as opposed to using confidence interval widths explicitly.

- Solving for n we get

$$n = \left[\frac{t_{n-1, 1-\alpha/2} \sigma}{\delta} \right]^2$$

- If n is large enough and $\alpha = 0.05$, required $n = 3.84 \left[\frac{\sigma}{\delta} \right]^2$
- Example: if want to be able to nail down μ to within $\pm 1\text{mmHg}$ when the patient standard deviation in blood pressure is 10mmHg , $n = 384$

```
sigma ← 10
delta ← 1
3.84 * (sigma / delta) ^ 2
```

[1] 384

- Advantages of planning for precision rather than power^q
 - do not need to select a single effect to detect
 - many studies are powered to detect a miracle and nothing less; if a miracle doesn't happen, the study provides **no** information
 - planning on the basis of precision will allow the resulting study to be interpreted if the P -value is large, because the confidence interval will not be so wide as to include both clinically significant improvement and clinically significant worsening

^qSee Borenstein M: *J Clin Epi* 1994; 47:1277-1285.

5.7

One Sample Method for a Probability

5.7.1

Frequentist Methods

ABD7

- Estimate a population probability p with a sample estimate \hat{p}
- Data: s “successes” out of n trials
- Maximum likelihood estimate of p is $\hat{p} = \frac{s}{n}$ (value of p making the data most likely to have been observed) = Bayesian posterior mode under a flat prior
- Approximate 2-sided test of $H_0 : p = p_0$ obtained by computing a z statistic
- A z -test is a test assuming that the *statistic* has a normal distribution; it is a t -test with infinite (∞) d.f.

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- The z -test follows the same general form as the t -test

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of numerator}}$$

- Example: $n = 10$ tosses of a coin, 8 heads; H_0 : coin is fair ($p = p_0 = \frac{1}{2}$)

$$z = \frac{.8 - .5}{\sqrt{(\frac{1}{2})(\frac{1}{2})/10}} = 1.897$$

- P -value = $2 \times$ area under a normal curve to the right of 1.897 = $2 \times 0.0289 = 0.058$ (this is also the area under the normal curve to the right of 1.897 + the area to the left of -1.897)

```

p  ← 0.8
p0 ← 0.5
n  ← 10
z  ← (p - p0) / sqrt(p0 * (1 - p0) / n)
c(z=z, Pvalue=2 * pnorm(-abs(z)))

```

z	Pvalue
1.89736660	0.05777957

- Approximate probability of getting 8 or more or 2 or fewer heads if the coin is fair is 0.058. **This is indirect evidence for fairness, is not the probability the null hypothesis is true, and invites the “absence of evidence is not evidence of absence” error.**
- Use exact methods if p or n is small

```
# Pr(X ≥ 8) = 1 - Pr(X < 8) = 1 - Pr(X ≤ 7)
pbinom(2, 10, 0.5) + 1 - pbinom(7, 10, 0.5)
```

```
[1] 0.109375
```

```
# Also compute as the probability of getting 0, 1, 2, 8, 9, 10 heads
sum(dbinom(c(0, 1, 2, 8, 9, 10), 10, 0.5))
```

```
[1] 0.109375
```

- Confidence interval for p
 - [Wilson's method](#) without continuity correction is recommended
 - for 8 of 10 heads here is the Wilson interval in addition to the exact binomial and normal approximation. The Wilson interval is the most accurate of the three.

```
binconf(8, 10, method='all')
```

	PointEst	Lower	Upper
Exact	0.8	0.4439045	0.9747893
Wilson	0.8	0.4901625	0.9433178
Asymptotic	0.8	0.5520820	1.0479180

5.7.2

Bayesian

- The single unknown probability p for binary Y -case is a situation where there is a top choice for prior distributions
- Number of events follows a binomial distribution with parameters p, n

- The beta distribution is for a variable having a range of $[0, 1]$, has two parameters α, β for flexibility, and is *conjugate* to the binomial distribution
 - The posterior distribution is simple: another beta distribution
- The mean of a beta-distributed variable is $\frac{\alpha}{\alpha+\beta}$ and its standard deviation is $\sqrt{\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)}}$
- Using a beta prior is equivalent to adding α successes and β failures to the data
- Posterior distribution of p is $\text{beta}(s + \alpha, n - s + \beta)$
- A uniform prior sets $\alpha = \beta = 1$
- In general an intuitive way to set the prior is to preset the mean then solve for the parameters that force $P(p > c) = a$ for given c and a
- For the 10 coin toss example let's set the prior mean of $P(\text{heads}) = \frac{1}{2}$ and $P(p > 0.8) = 0.05$, i.e. only a 0.05 chance that the probability of heads exceeds 0.8. Here are α and β satisfying these requirements:

```
# alpha/(alpha + beta) = 0.5 -> alpha=beta
alphas <- seq(0, 20, length=100000)
exceedanceProb <- 1 - pbeta(0.8, alphas, alphas)
alpha <- alphas[which.min(abs(exceedanceProb - 0.05))]
beta <- alpha
alpha.post <- 8 + alpha
beta.post <- 10 - 8 + beta
```

- The solution is $\alpha = \beta = 3.26$
- With the data $s = 8$ out of $n = 10$ the posterior distribution is $\text{beta}(11.26, 5.26)$

```
p <- seq(0, 1, length=300)
prior <- dbeta(p, alpha, beta)
post <- dbeta(p, alpha.post, beta.post)
curves <- list(Prior=list(x=p, y=prior),
                 Posterior=list(x=p, y=post))
labcurve(curves, pl=TRUE, xlab='p', ylab='Probability Density')
```

- From the posterior distribution we can get the credible interval, the probability that the probability of heads exceeds $\frac{1}{2}$, and the probability that the probability of heads is within ± 0.05 of fairness:

```
qbeta(c(.025, .975), alpha.post, beta.post)
```

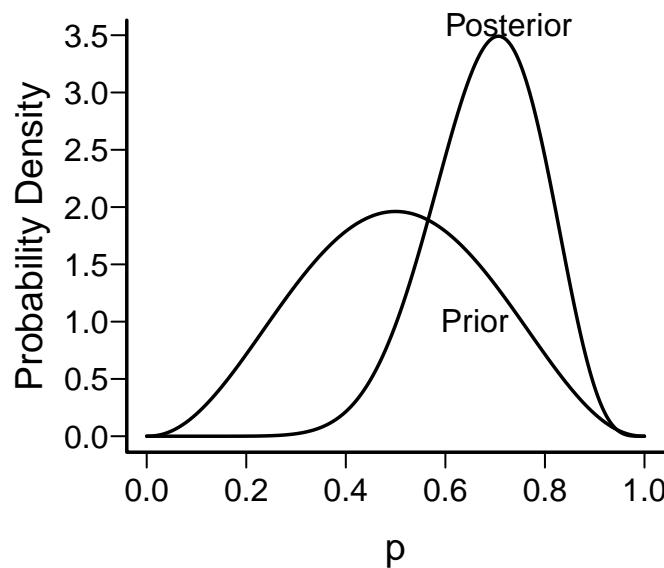


Figure 5.5: Prior and posterior distribution for the unknown probability of heads. The posterior is based on tossing 8 heads out of 10 tries.

```
[1] 0.4464199 0.8751905
```

```
1 - pbeta(0.5, alpha.post, beta.post)
```

```
[1] 0.9374297
```

```
pbeta(0.55, alpha.post, beta.post) - pbeta(0.45, alpha.post, beta.post)
```

```
[1] 0.100748
```

- Unlike the frequentist analysis, these are direct measures that are easier to interpret
- Instead of just providing evidence against a straw-man assertion, Bayesian posterior probabilities measure evidence in favor (as well as against) all possible assertions

5.7.3

Power and Sample Size

- Power \uparrow as $n \uparrow$, p departs from p_0 , or p_0 departs from $\frac{1}{2}$
- $n \downarrow$ as required power \downarrow or p departs from p_0

5.7.4

Sample Size for Given Precision

- Approximate 0.95 CL: $\hat{p} \pm 1.96\sqrt{\hat{p}(1 - \hat{p})/n}$
- Assuming p is between 0.3 and 0.8, it would not be far off to use the worst case standard error $\sqrt{1/(4n)}$ when planning
- n to achieve a margin of error δ in estimating p :

$$n = \frac{1}{4} \left[\frac{1.96}{\delta} \right]^2 = \frac{0.96}{\delta^2}$$

- Example: $\delta = .1 \rightarrow n = 96$ to achieve a margin of error of ± 0.1 with 0.95 confidence

```
nprec <- function(delta) round(0.25 * (qnorm(0.975) / delta) ^ 2)
nprec(0.1)
```

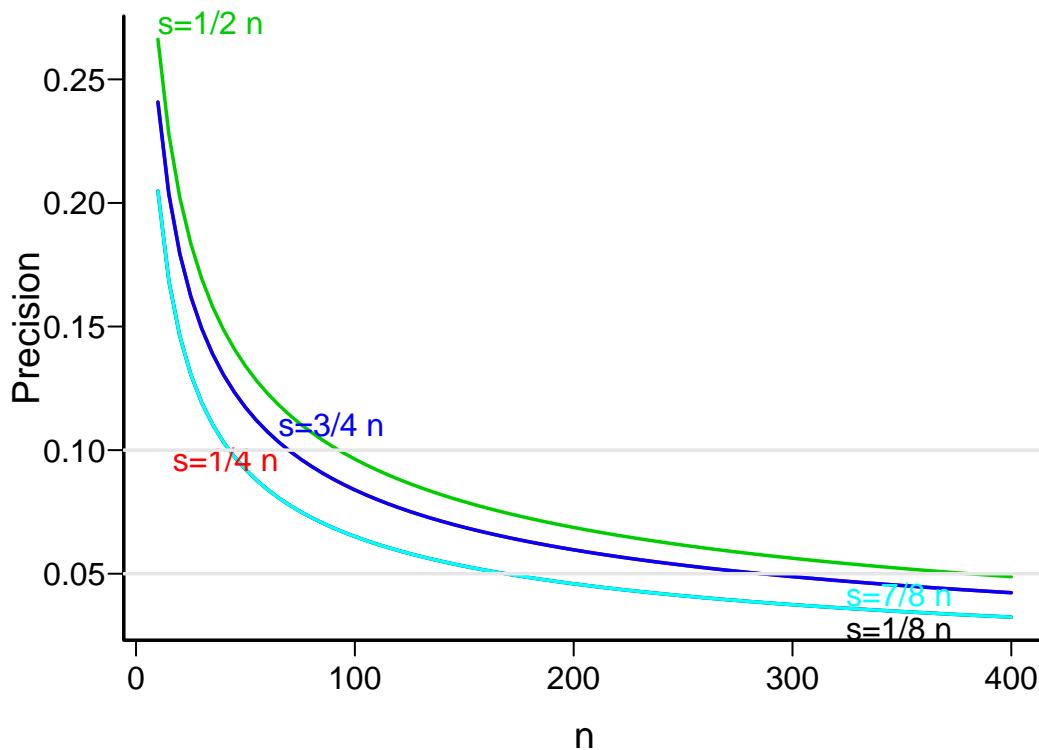
```
[1] 96
```

To achieve a margin of error of ± 0.05 even in the worst case where $p = 0.5$ one needs $n = 384$.

For Bayesian precision calculations we can solve n that achieves a given width of, say, a 0.95 credible interval:

- Use a flat beta prior, i.e., with $\alpha = \beta = 1$
- Posterior distribution for p is $\text{beta}(s + 1, n - s + 1)$
- Compute CI half-widths for varying n for selected values of s

```
n <- seq(10, 400, by=5)
k <- c(1/8, 1/4, 1/2, 3/4, 7/8)
ck <- paste0('s=', c('1/8', '1/4', '1/2', '3/4', '7/8'), ', n')
r <- list()
for(i in 1 : 5) {
  ciu <- qbeta(0.975, k[i] * n + 1, n - k[i] * n + 1)
  cil <- qbeta(0.025, k[i] * n + 1, n - k[i] * n + 1)
  r[[ck[i]]] <- list(x=n, y=(ciu - cil) / 2)
}
labcurve(r, xlab='n', ylab='Precision', col=1:5, pl=TRUE)
abline(h=c(0.05, 0.1), col=gray(0.9))
```

Figure 5.6: Half-widths of 0.95 credible intervals for p using a flat prior

- As with confidence intervals, precision is worst when $\frac{1}{2}$ of observations are successes, so often best to plan on worst case
- Same sample size needed as with frequentist (since prior is flat)
- Easy to modify for other priors

To put this in the context of relative errors, suppose that one wants to estimate the odds that an event will occur, to within a certain multiplicative margin of error (MMOE) with 0.95 confidence using frequentist methods. What is the MMOE as a function of the unknown p when $n = 384$? The standard error of the log odds is approximately $\sqrt{\frac{1}{np(1-p)}}$, and the half-width of a 0.95 confidence interval for the log odds is approximately 1.96 times that. Fix $n = 384$ and vary p to get the MMOE that is associated with the same sample size as a universal absolute margin of error of 0.05.

```
p <- seq(0.01, 0.99, length=200)
mmoe <- exp(1.96 / sqrt(384 * p * (1 - p)))
plot(p, mmoe, type='l', xlab='Unknown Probability p', ylab='MMOE')
minor.tick()
```

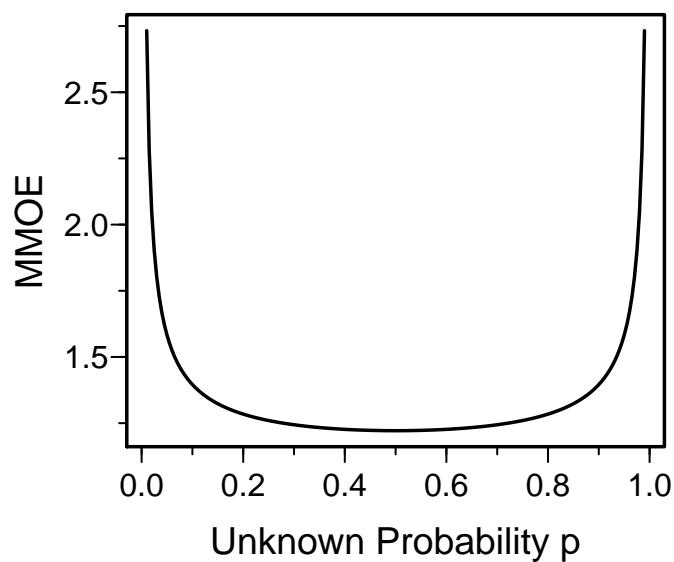


Figure 5.7: Multiplicative margin of error in estimating odds when $n = 384$ and the margin of error in estimating the absolute probability is ≤ 0.05 .

5.8

Paired Data and One-Sample Tests

ABD11

- To investigate the relationship between smoking and bone mineral density, Rosner presented a paired analysis in which each person had a nearly perfect control which was his or her twin
- Data were normalized by dividing differences by the mean density in the twin pair (need to check if this normalization worked)
- Note:** It is almost never appropriate to compute mean percent change (a 100% increase is balanced by a 50% decrease) but we plunge ahead anyway
- Computed density in heavier smoking twin minus density in lighter smoking one
- Mean difference was -5% with $se=2.0\%$ on $n = 41$
- The t statistic we've been using works here, once within-pair differences are formed
- H_0 : mean difference between twins is zero ($\mu_0 = 0$)

$$t_{40} = \frac{\bar{x} - \mu_0}{se} = -2.5 \\ P = 0.0166$$

```
xbar  ← -5
se    ← 2
n     ← 41
mu0  ← 0
tstat ← (xbar - mu0) / se
pval  ← 2 * (1 - pt(abs(tstat), n - 1))
c(tstat=tstat, Pvalue=pval)
```

tstat	Pvalue
-2.50000000	0.01662035

5.9

Two Sample Test for Means

ABD12
 

- Two groups of different patients (unpaired data)
- Much more common than one-sample tests
- As before we are dealing for now with parametric tests assuming the raw data arise from a normal distribution
- We assume for now that the two groups have the same spread or variability in the distributions of responses^r

5.9.1

Frequentist t -Test

- Test whether population 1 has the same mean as population 2
- Example: pop. 1=all patients with a certain disease if given the new drug, pop. 2=standard drug
- $H_0 : \mu_1 = \mu_2$ (this can be generalized to test $\mu_1 = \mu_2 + \delta$, i.e., $\mu_1 - \mu_2 = \delta$). The *quantity of interest* or *QOI* is $\mu_1 - \mu_2$
- 2 samples, of sizes n_1 and n_2 from two populations
- Two-sample (unpaired) t -test assuming normality and equal variances—recall that if we are testing against an H_0 of **no effect**, the form of the t test is

$$t = \frac{\text{point estimate of QOI}}{\text{se of numerator}}$$

- Point estimate QOI is $\bar{x}_1 - \bar{x}_2$
- As with 1-sample t -test the difference in the numerator is judged with respect to

^rRosner covers the unequal variance case very well. As nonparametric tests have advantages for comparing two groups and are less sensitive to the equal spread assumption, we will not cover the unequal variance case here.

the precision in the denominator (combination of sample size and subject-to-subject variability); like a signal:noise ratio

- Variance of the sum or difference of two independent means is the sum of the variance of the individual means
- This is $\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2[\frac{1}{n_1} + \frac{1}{n_2}]$
- Need to estimate the single σ^2 from the two samples
- We use a weighted average of the two sample variances:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- True standard error of the difference in sample means: $\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- Estimate: $s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, so
$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$
- d.f. is the sum of the individual d.f., $n_1 + n_2 - 2$, where the -2 is from our having to estimate the center of two distributions
- If H_0 is true t has the $t_{n_1+n_2-2}$ distribution
- To get a 2-tailed P -value we compute the probability that a value from such a distribution is farther out in the tails of the distribution than the observed t value is (we ignore the sign of t for a 2-tailed test)
- Example: $n_1 = 8, n_2 = 21, s_1 = 15.34, s_2 = 18.23, \bar{x}_1 = 132.86, \bar{x}_2 = 127.44$

$$\begin{aligned}s^2 &= \frac{7(15.34)^2 + 20(18.23)^2}{7 + 20} = 307.18 \\ s &= \sqrt{307.18} = 17.527 \\ se &= 17.527 \sqrt{\frac{1}{8} + \frac{1}{21}} = 7.282 \\ t &= \frac{5.42}{7.282} = 0.74\end{aligned}$$

on 27 d.f.

- $P = 0.463$ (see R code below)
- Chance of getting a difference in means as larger or larger than 5.42 if the two populations really have the same means is 0.463
- → little evidence for concluding the population means are different

```
n1      ← 8;          n2 ← 21
xbar1 ← 132.86; xbar2 ← 127.44
s1      ← 15.34;     s2 ← 18.23
s      ← sqrt(((n1 - 1) * s1 ^ 2 + (n2 - 1) * s2 ^ 2) / (n1 + n2 - 2))
se      ← s * sqrt(1 / n1 + 1 / n2)
tstat ← (xbar1 - xbar2) / se
pval ← 2 * (pt(- abs(tstat), n1 + n2 - 2))
c(s=s, se=se, tstat=tstat, Pvalue=pval)
```

s	se	tstat	Pvalue
17.5265589	7.2818380	0.7443176	0.4631137

5.9.2

Confidence Interval

Assuming equal variances

$$\bar{x}_1 - \bar{x}_2 \pm t_{n_1+n_2-2, 1-\alpha/2} \times s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is a $1 - \alpha$ CL for $\mu_1 - \mu_2$, where s is the pooled estimate of σ , i.e., $s\sqrt{\dots}$ is the estimate of the standard error of $\bar{x}_1 - \bar{x}_2$

5.9.3

Bayesian t -Test

- As with one-sample Bayesian t -test we relax the normality assumption by using a t distribution for the raw data (more robust analysis)
- A linear model for the two-sample t test is

$$Y = \mu_0 + \delta[\text{group B}] + \epsilon$$

where μ_0 (the intercept) is the unknown group A mean, δ is the B-A difference in means, and ϵ is the irreducible residual (assuming we have no covariates to adjust for)

- Assume:

- ϵ has a t -distribution with ν d.f.
- ν has a prior that allows the data distribution to be anywhere from heavy-tailed to normal
- μ_0 has a fairly wide prior distribution (no prior knowledge may be encoded by using a flat prior)
- δ has either a prior that is informed by prior reliable research or biological knowledge, or has a skeptical prior
- residual variance σ^2 is allowed to be different for groups A and B, with a normal prior on the log of the variance ratio that favors equal variance but allows the ratio to be different from 1 (but not arbitrarily different)

Note: by specifying independent priors for μ_0 and δ we

- induce correlations in priors for the two means
- assume we know more about δ than about the individual true per-group means

To specify the SD for the prior for the log variance ratio:

- Let r be unknown ratio of variances
- Assume $P(r > 1.5) = P(r < \frac{1}{1.5}) = \gamma$
- The required SD is $\frac{\log 1.5}{-\Phi^{-1}(\gamma)}$
- For $\gamma = 0.15$ the required SD is:

```
log(1.5) / -qnorm(0.15)
```

```
[1] 0.3912119
```

We are assuming a mean of zero for $\log(r)$ so we favor $r = 1$ and give equal chances to ratios smaller or larger than 1.

5.9.4

Power and Sample Size

- Consider the frequentist model

- Power increases when

- $\Delta = |\mu_1 - \mu_2| \uparrow$

- $n_1 \uparrow$ or $n_2 \uparrow$

- n_1 and n_2 are close

- $\sigma \downarrow$

- $\alpha \uparrow$

- Power depends on $n_1, n_2, \mu_1, \mu_2, \sigma$ approximately through

$$\frac{\Delta}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Note that when computing power using a program that asks for μ_1 and μ_2 you can just enter 0 for μ_1 and enter Δ for μ_2 , as only the difference matters

- Often we estimate σ from pilot data, and to be honest we should make adjustments for having to estimate σ although we usually run out of gas at this point (Bayes would help)

- Use the R `pwr` package, or the power calculator at statpages.org/#Power or PS

- Example:

Get a pooled estimate of σ using s above (17.52656)

Use $\Delta = 5, n_1 = n_2 = 100, \alpha = 0.05$

```
delta <- 5
require(pwr)
pwr.t2n.test(n1=100, n2=100, d=delta / s, sig.level = 0.05)
```

```
t test power calculation

n1 = 100
n2 = 100
d = 0.2852813
sig.level = 0.05
power = 0.5189751
alternative = two.sided
```

- Sample size depends on $k = \frac{n_2}{n_1}$, Δ , power, and α
- Sample size \downarrow when
 - $\Delta \uparrow$
 - $k \rightarrow 1.0$
 - $\sigma \downarrow$
 - $\alpha \uparrow$
 - required power \downarrow
- An approximate formula for required sample sizes to achieve power = 0.9 with $\alpha = 0.05$ is

$$n_1 = \frac{10.51\sigma^2(1 + \frac{1}{k})}{\Delta^2}$$

$$n_2 = \frac{10.51\sigma^2(1 + k)}{\Delta^2}$$

- Exact calculations assuming normality

```
pwr.t.test(d = delta / s, sig.level = 0.05, power = 0.8)
```

```
Two-sample t test power calculation

n = 193.8463
d = 0.2852813
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
```

- If used same total sample size of 388 but did a 2:1 randomization ratio to get 129 in one group and 259 in the other, the power is less

```
pwr.t2n.test(n1 = 129, n2 = 259, d = delta / s, sig.level = 0.05)
```

```
t test power calculation

n1 = 129
n2 = 259
d = 0.2852813
sig.level = 0.05
power = 0.7519836
alternative = two.sided
```

What is the difference in means that would yield a 2-sided P -value of exactly 0.05 for a two-sample t -test with normality and equal variances when the sample sizes are both equal to $\frac{n}{2}$? We solve for $\hat{\Delta} = \bar{x}_1 - \bar{x}_2$ such that $t_{n-2,1-\alpha/2} = \frac{\hat{\Delta}}{2s/\sqrt{n}}$, giving

$$\hat{\Delta} = \frac{2 \times t_{n-2,1-\alpha/2} \times s}{\sqrt{n}}$$

For total sample sizes of 10, 50, and 100, the “magic” values of the observed difference are the following multiples of the observed standard deviation s :

```
n <- c(10, 50, 100)
tcrit <- qt(0.975, n-2)
2 * tcrit / sqrt(n)
```

```
[1] 1.4584451 0.5686934 0.3968935
```

Note that these thresholds are independent of the power and the effect size used in the power calculation.

5.9.5

Sample Size for a Given Precision

To design a study that will nail down the estimate of $\mu_1 - \mu_2$ to within $\pm\delta$ with $1 - \alpha$ confidence when $n_1 = n_2 = n$, and when n is large enough so that the critical value $t_{2n-2,1-\alpha/2}$ may be approximated by the critical value from the normal distribution, say z ($z = 1.96$ when $\alpha = 0.05$):

$$n = 2 \left[\frac{z\sigma}{\delta} \right]^2$$

When $\alpha = 0.05$, $n = 7.68 \left[\frac{\sigma}{\delta} \right]^2$

5.9.6

Equating Margin of Error to Detectable Difference

Suppose that a two-arm study is designed to detect a difference Δ in two means with power 0.9 at the $\alpha = 0.05$ level. For large enough sample sizes, the margin of error for estimating the true difference in means for that study will be $\delta = \Delta \sqrt{\frac{7.68}{21.02}} = 0.604\Delta$.

5.9.7

Checking Assumptions of the *t*-test

- Comprehensive assessment of all assumptions except independence of observations:
 - Compute the two empirical cumulative distribution functions
 - Transform each using the inverse normal z transformation
 - See if both curves are linear (checks normality assumption) and parallel (equal variance assumption)^s
- Box plot (one box for each of 2 groups): look for equal spread (IQR)
- Informally compare s_1 and s_2 ^t
- With the Bayesian *t*-test the only important assumption to check is symmetry of the data distribution^u

^sThere are formal tests of normality but in smaller samples these have insufficient power to detect important non-normality.

^tRosner 8.6 shows how to make formal comparisons, but beware that the variance ratio test depends on normality, and it may not have sufficient power to detect important differences in variances.

^uSince we are allowing for heavier tails than the Gaussian distribution by using a *t* distribution for the raw data

5.10**Comprehensive Example: Two sample *t*-test****5.10.1****Study Description**

From bit.ly/data-t2

- Assess the effect of caffeine (vs. placebo) on muscle metabolism, measured by the respiratory exchange ratio (RER; ratio of CO₂ produced to O₂ consumed)
- Treatment was randomized to 18 subjects; parallel group RCT
- Goal: study effect of caffeine on RER
- Must take log of RER to have a symmetric measure
 - μ_0 = mean log RER for placebo
 - μ_1 = mean log RER for caffeine
 - Fold-change effect: $\exp(\mu_1 - \mu_0)$
 - Estimate $\mu_1 - \mu_0$
 - $H_0 : \mu_0 = \mu_1$
 - $H_1 : \mu_0 \neq \mu_1$
- Note: a good statistician will take such ratios with a grain of salt; need to verify that the meaning of the ratio is independent of O₂

5.10.2

Power and Sample Size

- Suppose that a pilot study or previous published research estimated $\sigma = 0.1$ for log RER
- Effect size Δ is on the log RER scale
- Anti-log to get effect in terms of fold change
- Determine the number of subjects needed (in each group) for several value of effect size Δ ($\Delta = |\mu_1 - \mu_0|$) in order to have 0.9 power with $\alpha = 0.05$

```
require(pwr)
s <- 0.1
fc <- c(1.1, 1.15, 1.2, 1.25, 1.5)
n <- integer(length(fc))
i <- 0
for(foldchange in fc) {
  i <- i + 1
  n[i] <- ceiling(pwr.t.test(d=log(foldchange) / s, power=0.9)$n)
}
data.frame('Fold Change'=fc, Delta=round(log(fc), 3), 'N per group'=n,
           check.names=FALSE)
```

	Fold Change	Delta	N per group
1	1.10	0.095	25
2	1.15	0.140	12
3	1.20	0.182	8
4	1.25	0.223	6
5	1.50	0.405	3

- If caffeine modifies RER by a factor of 1.15, by enrolling 12 subjects in each group we will have 0.9 power to detect an effect
- For $n = 12$ per group the margin of error for estimating Δ at the 0.95 level is given below
See Section 5.9.5
- This is anti-logged to obtain the multiplicative margin of error for estimating the caffeine:placebo ratio of RERs

```
z <- qnorm(0.975); n <- 12
# This is approximate; use z <- qt(0.975, 12+12-2) for accuracy
```

```
moe ← z * s * sqrt(2 / n)
mmoe ← exp(moe)
c('Margin of Error'=moe, 'Multiplicative Margin of Error'=mmoe)
```

Margin of Error	Multiplicative Margin of Error
0.08001519	1.08330353

5.10.3

Collected Data

```
tx ← factor(c(rep('placebo', 9), rep('caffeine', 9)), c('placebo', 'caffeine'))
rer ← c(105, 119, 100, 97, 96, 101, 94, 95, 98,
       96, 99, 94, 89, 96, 93, 88, 105, 88) / 100
d ← data.frame(subject=1:18, tx, rer, logrer=log(rer))
print(d, digits=3, row.names=FALSE)
```

subject	tx	rer	logrer
1	placebo	1.05	0.04879
2	placebo	1.19	0.17395
3	placebo	1.00	0.00000
4	placebo	0.97	-0.03046
5	placebo	0.96	-0.04082
6	placebo	1.01	0.00995
7	placebo	0.94	-0.06188
8	placebo	0.95	-0.05129
9	placebo	0.98	-0.02020
10	caffeine	0.96	-0.04082
11	caffeine	0.99	-0.01005
12	caffeine	0.94	-0.06188
13	caffeine	0.89	-0.11653
14	caffeine	0.96	-0.04082
15	caffeine	0.93	-0.07257
16	caffeine	0.88	-0.12783
17	caffeine	1.05	0.04879
18	caffeine	0.88	-0.12783

```
require(ggplot2) # Fig 5.8
ggplot(d, aes(x=tx, y=rer, group=tx)) +
  geom_boxplot(col='lightyellow1', alpha=.3, width=.3) +
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
  stat_summary(fun.y=median, geom="point", col='red', shape=5, size=3) +
  xlab('') + ylab('RER') + coord_flip()
```

5.10.4

Frequentist *t*-Test

To demonstrate difficulties in checking model assumptions with small n , consider the comprehensive approach by checking for linearity and parallelism (if equal variance

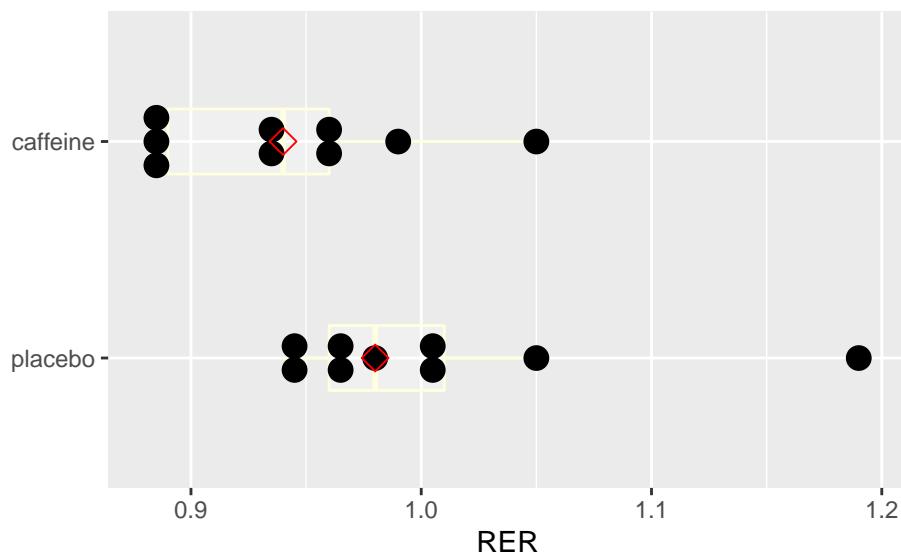


Figure 5.8: Data for two-sample RCT for effect of caffeine on respiratory exchange ratio. Diamonds depict medians.

assumption is used) of z -transformed empirical cumulative distributions.

```
Ecdf(~ log(rer), groups=tx, fun=qnorm, data=d,
    xlab='log(RER)', ylab='Inverse Normal ECDF') # Fig. 5.9
```

It is more accepted in practice to now use the form of the t -test that does not assume equal variances in the two independent groups. The unequal-variance t -test is used here. Note that to compute a decent approximation to the P -value requires the use of a “trick” d.f. when looking up against a t distribution.

```
ttest ← t.test(log(rer) ~ tx, data=d)
# Note that for the CI t.test is using caffeine as the reference group
ttest
```

```

Welch Two Sample t-test

data: log(rer) by tx
t = 2.0622, df = 15.337, p-value = 0.05655
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.002027639 0.130381419
sample estimates:
mean in group placebo mean in group caffeine
0.003115688          -0.061061202
```

- Interpretation

- Subjects given caffeine have on average a log RER that is 0.064 lower (0.95 CI: [-0.13, 0.002]) than individuals given placebo

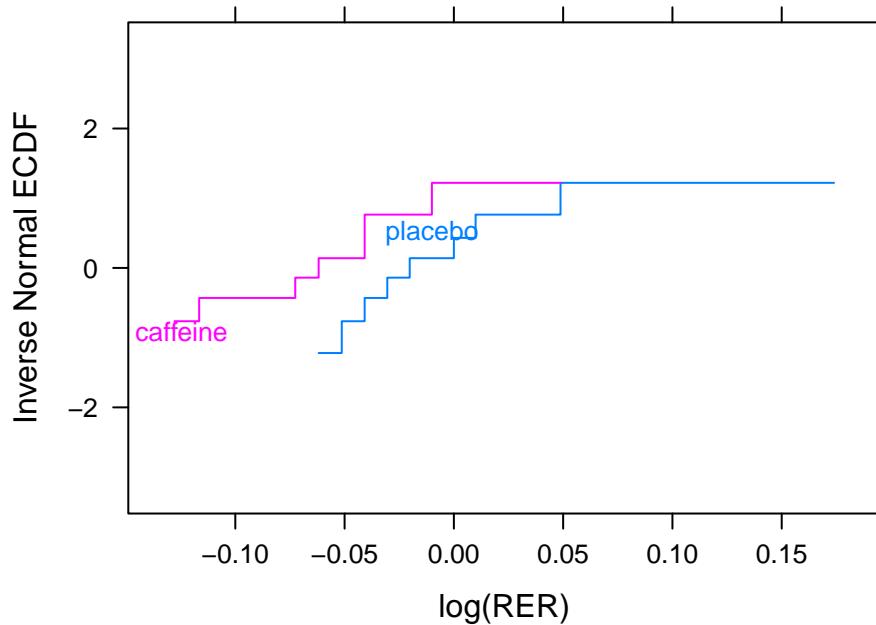


Figure 5.9: Stratified empirical cumulative distribution functions by treatment for checking all assumptions of the two-sample t -test. ECDFs are inverse normal transformed.

- By anti-logging these 3 numbers we get the fold change scale:
 - * Median^v fold change is 0.94
 - * 0.95 CI for caffeine:placebo fold change: [0.878, 1.002]

5.10.5

Bayesian t -Test

The R `brms` package makes it easy to specify an unequal variance model, because it allows one to specify a separate model for the log of the standard deviation. $\log(\sigma)$ can even depend on continuous covariates!^w `brms` models standard deviation parameters on the log scale. As before, analysis is on the $\log(\text{RER})$ scale.

```
require(brms)
# set priors

# flat (non-informative) prior for intercept
pr0 ← set_prior("", class="Intercept")
```

^vIf the log ratio has a normal distribution, the log ratio has the same mean as the median and its median anti-logged value is the anti-log of the mean (=median) log difference. The mean on the anti-logged scale is a more complex function that involves SD of log RER.

^wThanks to Nathan James of the Vanderbilt Department of Biostatistics for providing R code for this section, and for providing the explanation for the output of the `prior_summary` function.

```
# normal(0,3) prior for difference in log ratios
pr1 <- set_prior("normal(0,0.25)", class="b", coef="txcaffeine")

# normal(0,0.3912) for log SD ratio
pr2 <- set_prior("normal(0,0.3912)", class="b", coef="txcaffeine",
                  dpar="sigma")

# fit model for two groups assuming Y follows a t distribution so less
# sensitive to outliers
# Each group has different mean and SD but same df
# The prior for the log SD for the reference group is scale t with 3 d.f.
# Prior for nu is gamma(2, 0.1)
# The bf function is used to create a compound R model formula, here
# for the mean model and the log sigma model

f <- brm(bf(log(rer) ~ tx, sigma ~ tx), data=d, family=student,
          prior=c(pr0,pr1,pr2), seed=1202)
```

There are 5 parameters in this model (2 for the regression on student-*t* mean, 2 for the regression on student-*t* scale (σ), and 1 for student-*t* d.f. ν). The output from `prior_summary()` shows all the parameters *and* parameter classes that can be assigned. Lines 1 & 3 in the output below are the ‘classes’ of all non-intercept regression coefficients for the student-*t* mean and student-*t* scale, respectively. When there are multiple coefficients it is often convenient to specify a prior for all parameters in a class rather than for each individual parameter. For example, in a model with 10 covariates using `set_prior("normal(0,1)", class="b")` would give each of the 10 corresponding coefficient parameters a standard normal prior. For our model the class priors are superseded by the individual parameter priors.

```
# Show priors in effect
prior_summary(f)
```

	prior	class	coef	group	resp	dpar	nlnpar	bound
1		b						
2	normal(0,0.25)	b	txcaffeine					
3		b				sigma		
4	normal(0,0.3912)	b	txcaffeine			sigma		
5		Intercept						
6	student_t(3, 0, 10)	Intercept				sigma		
7	gamma(2, 0.1)		nu					

```
# model summary
# Note: sigma is on log scale, use exp(sigma) to get to natural scale
# Intercept is mean for placebo
# sigma_Intercept is log(sd) for placebo
# txcaffeine is change in mean for caffeine compared to placebo
# sigma_txcffeine is change in log(sd) for caffeine compared to placebo
f
```

Family: student
Links: mu = identity; sigma = log; nu = identity

```

Formula: log(rer) ~ tx
         sigma ~ tx
Data: d (Number of observations: 18)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

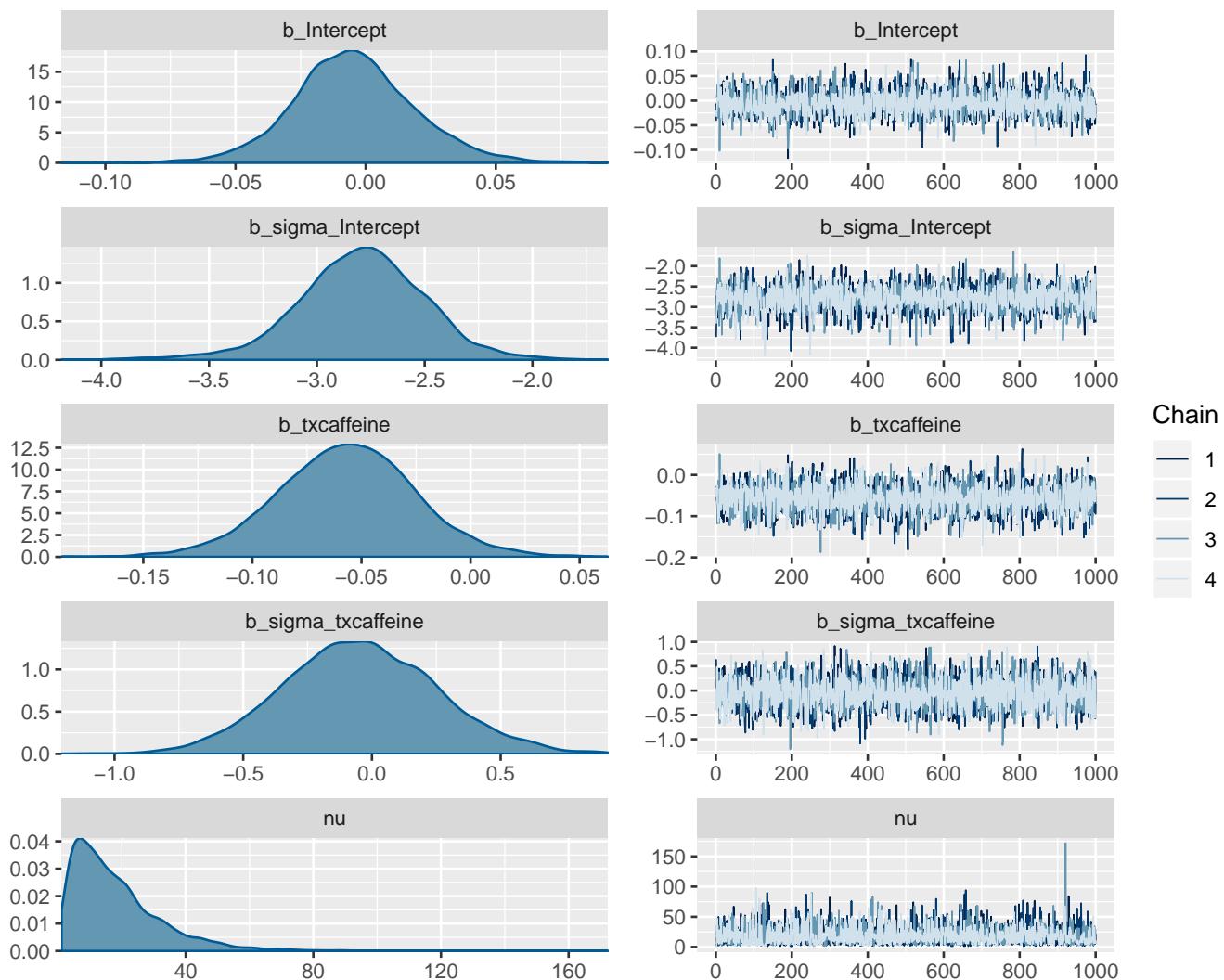
Population-Level Effects:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     -0.00      0.02    -0.05     0.04 1.00    3747    2751
sigma_Intercept -2.79      0.30   -3.43    -2.22 1.00    2642    2230
txcaffeine     -0.06      0.03   -0.12     0.01 1.00    4127    2533
sigma_txcffeine -0.05      0.30   -0.63     0.57 1.00    4114    2775

Family Specific Parameters:
Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
nu       18.17     14.05    2.28    53.10 1.00    2671    1753

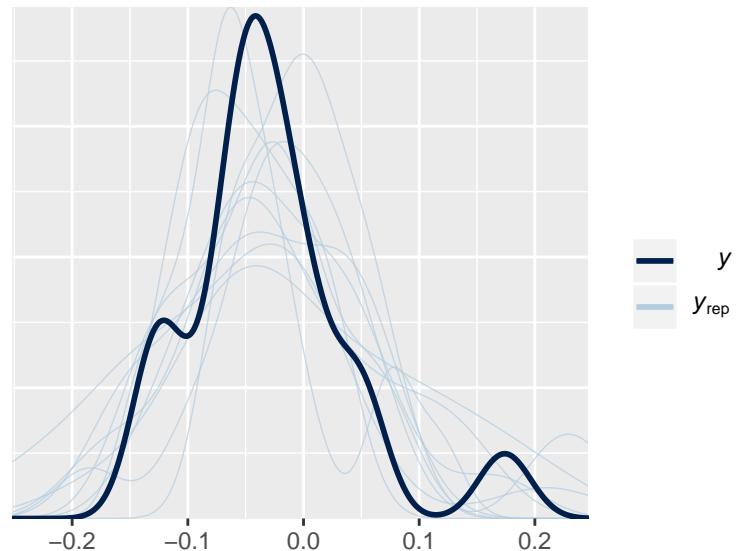
Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

```

```
# plot kernel density est and trace plots for posterior parameters
plot(f)
```

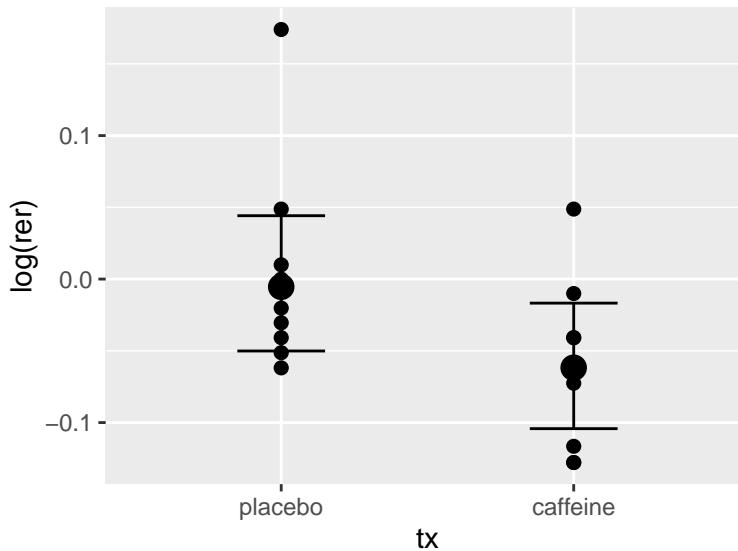


```
# posterior predictive check
pp_check(f)
```



```
# marginal effects
```

```
plot(marginal_effects(f), points = TRUE)
```



```
# posterior parameter samples
p <- as.data.frame(f)
meanplacebo <- p[, 'b_Intercept']
delta <- p[, 'b_txcaffeine']
sdplacebo <- exp(p[, 'b_sigma_Intercept'])
sdratio <- exp(p[, 'b_sigma_txcaffeine'])
nu <- p[, 'nu']

# histogram of posterior distribution of difference in mean log RER
hist(delta, nclass=50, main='')

# Posterior density for caffeine:placebo fold change in RER
plot(density(exp(delta)), xlab='Fold Change in RER', main='')
abline(v=1, col=gray(0.85))

# Posterior density of SD ratio
plot(density(sdratio), main='', xlab='SD Ratio')
abline(v=1, col=gray(0.85))

# Posterior prob that difference in means < 0
# Recall that the P operator was defined previously
# (based on mean of logical or 0/1 values = proportion positive =
# posterior probability to within simulation error
# This is the same as P(fold change < 1)
P(delta < 0)
```

```
[1] 0.965
```

```
P(exp(delta) < 1)
```

```
[1] 0.965
```

```
# Prob that caffeine results in a physiologically noticeable response
P(exp(delta) < 0.95)
```

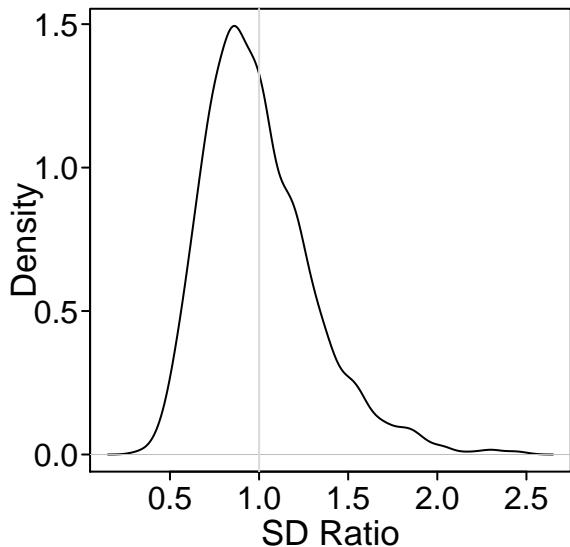
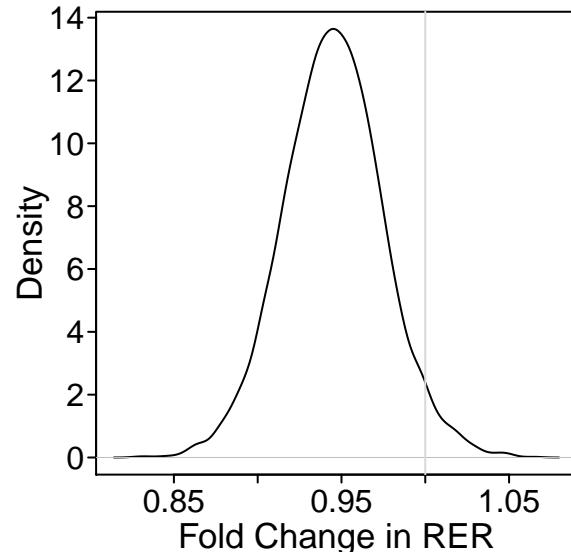
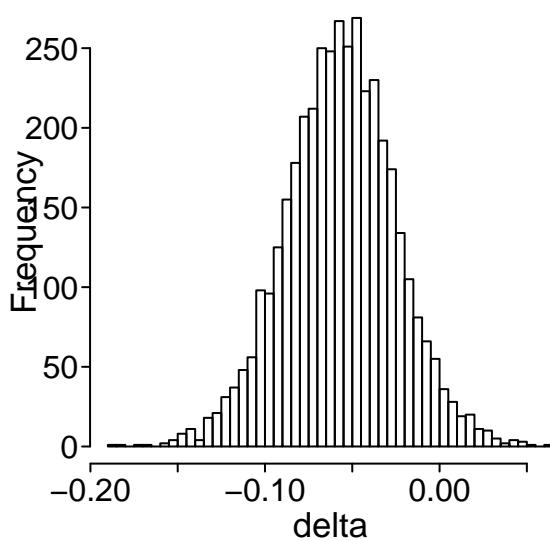
```
[1] 0.56825
```

```
# Prob that caffeine and placebo have similar response
P(exp(delta) > 0.975 & exp(delta) < 1/0.975)
```

```
[1] 0.141
```

```
# Compute posterior probability of approximate normality
P(nu > 20)
```

```
[1] 0.3535
```



Note that the 2-tailed P -value of 0.057 may tempt bright-line threshold advocates to conclude nothing more than insufficient evidence to reject the assumption that caffeine does not modify RER (or worse yet to just declare an “insignificant” result or even worse that caffeine does not modify RER). The Bayesian result shows that under a fairly skeptical prior for log RER one would do well to play the odds in betting on caffeine having an effect.

There is little evidence to support an assumption of normality of log RER within a treatment group.

Now compare other results results with the frequentist analysis.

```
means <- with(d, tapply(logrer, tx, mean))
sds <- with(d, tapply(logrer, tx, sd))
a <- c(means[1], pmode(meanplacebo), mean(meanplacebo), median(meanplacebo))
b <- c(sds[1], pmode(sdplacebo), mean(sdplacebo), median(sdplacebo))
w <- c(diff(means), pmode(delta), mean(delta), median(delta))
z <- c(sds[2] / sds[1], pmode(sdratio), mean(sdratio), median(sdratio))
x <- rbind(a, b, w, z)
colnames(x) <- c('Sample', 'Posterior Mode', 'Posterior Mean',
                  'Posterior Median')
rownames(x) <- c('Placebo mean', 'Placebo SD', 'Caffeine - Placebo Mean',
                  'Caffeine / Placebo SD')
round(x, 2)
```

	Sample	Posterior Mode	Posterior Mean	Posterior Median
Placebo mean	0.00	-0.01	0.00	-0.01
Placebo SD	0.07	0.06	0.06	0.06
Caffeine - Placebo Mean	-0.06	-0.06	-0.06	-0.06
Caffeine / Placebo SD	0.81	0.86	1.00	0.95

```
# 0.95 credible interval for delta
quantile(delta, c(0.025, .975))
```

```
2.5%          97.5%
-0.120428468  0.005351076
```

```
# 0.95 confidence interval for delta
rev(- ttest$conf.int)  # negate since t.test used different reference
```

```
[1] -0.130381419  0.002027639
```

For an excellent tutorial on the use of `brms` for a two-sample *t*-test see bit.ly/brms-t by Matti Vuorre

5.11

The Problem with Hypothesis Tests and P -values Revisited

5.11.1

Hypothesis Testing

- Existence of ESP is a hypothesis
- Assessing effects of drugs, procedures, devices involves estimation
- Many studies powered to detect huge effect
- If effect is not huge, no information from study

5.11.2

P -Values

ABD6.2

- Only provide evidence against a *null* hypothesis, **never** evidence for something
- Probability of a statistic *more* impressive as yours **if** H_0 true
- Not a probability of an effect or difference (same problem with sensitivity and specificity)
- **No** conclusion possible from large P -values
- Cannot conclude clinical relevance from small P
- Adjustment of P -values for multiple tests is controversial and there is insufficient consensus on how to choose an adjustment method
- Declaring a result as “significant” or “non-significant” is completely arbitrary and has come to mean almost nothing

- They rely on completely arbitrary P -value cutoffs such as 0.05
- American Statistical Association is on record recommending not using any cutoff or words like *significant*: bit.ly/asa-p bit.ly/asa-p2

5.11.3

How Not to Present Results

ABD6.2

- $P = 0.02$ — let's put this into clinical practice ignoring the drug's cost or clinical effectiveness
- $P = 0.4$ — this drug does not kill people
- $P = 0.2$ but there is a trend in favor of our blockbuster drug
- The observed difference was 6mmHg and we rejected H_0 so the true effect is 6mmHg.
- The proportion of patients having adverse events was 0.01 and 0.03; the study wasn't powered to detect adverse event differences so we present no statistical analysis
- The reduction in blood pressure was 6mmHg with 0.95 C.L. of [1mmHg, 11mmHg]; the drug is just as likely to only reduce blood pressure by 1mmHg as it is by 6mmHg.
- The serum pH for the 15 dogs was 7.3 ± 0.1 (mean \pm SE instead of SD or IQR)

5.11.4

How to Present Results

ABD6.2

- Estimates should be accompanied by uncertainty intervals or posterior distributions
- Confidence limits can be computed without regard to sample size or power
- A computed value from a sample is only an estimate of the population value, whether or not you reject H_0

- Best to think of an estimate from a study as a fuzz, not a point
- To present variability of subjects, use SD or IQR, **not** SE (SE is the precision of the *mean* of subjects)
- If you must use *P*-values, provide the *P*-value to 3 significant digits and don't declare results as *significant* or *no significant difference*
- See <http://bit.ly/datamethods-freq-results> for some guidelines for presenting frequentist results, and fharrell.com/post/bayes-freq-stmts for examples of Bayesian vs. frequentist summaries

5.12

Study Design Considerations



The majority of studies phrased as hypothesis testing experiments are actually estimation studies, so it is usually preferred to base sample size justifications on precision (margin of error). Whether using effect sizes in power calculations or margins of error in precision calculations, the quantity of interest should be taken on the original dependent variable scale or a transformation of it such as odds or hazard.

5.12.1

Sizing a Pilot Study

Frequently, pilot studies are used to obtain estimates of variability that allow the sample size to be calculated for a full study. With a continuous response variable, one can think of the adequacy of the sample size in terms of the fold change or multiplicative margin of error (MMOE) in the estimate s of the population standard deviation σ .

When a sample of size n is drawn from a normal distribution, a $1 - \alpha$ two-sided confidence interval for the unknown population variance σ^2 is given by

$$\frac{n-1}{\chi_{1-\alpha/2,n-1}^2} s^2 < \sigma^2 < \frac{n-1}{\chi_{\alpha/2,n-1}^2} s^2, \quad (5.1)$$

where s^2 is the sample variance and $\chi_{\alpha/2,n-1}^2$ is the α critical value of the χ^2 distribution with $n - 1$ degrees of freedom. The MMOE for estimating σ is

$$\sqrt{\max\left(\frac{\chi_{1-\alpha/2,n-1}^2}{n-1}, \frac{n-1}{\chi_{\alpha/2,n-1}^2}\right)} \quad (5.2)$$

```

n      ← 10:300
low   ← sqrt((n - 1) / qchisq(.975, n - 1))
hi    ← sqrt((n - 1) / qchisq(.025, n - 1))
m     ← pmax(1 / low, hi)
ggplot(data.frame(n, m), aes(x=n, y=m)) + geom_line() +
  ylab('MMOE for s')
nmin ← min(n[m ≤ 1.2])

```

From the above calculations, to achieve a MMOE of no worse than 1.2 with 0.95 confidence when estimating σ requires a sample size of 70 subjects. A pilot study with

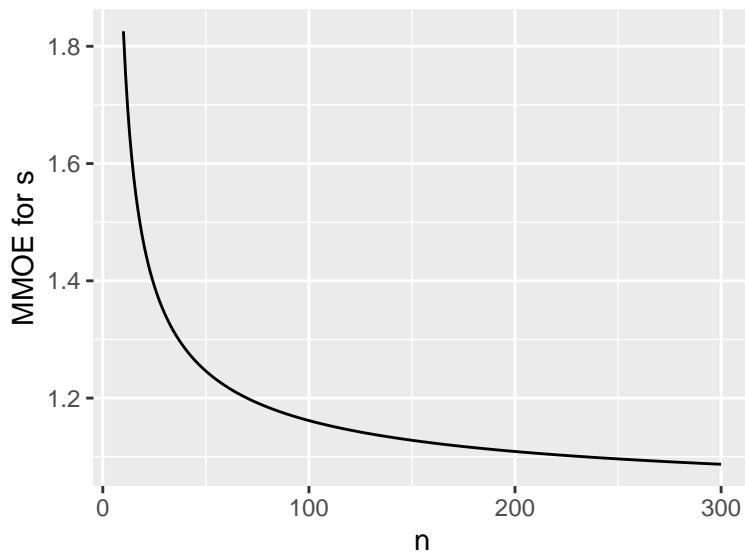


Figure 5.10: Multiplicative margin of error in estimating σ as a function of sample size, with 0.95 confidence

$n = 20$ will achieve a MMOE of 1.46 in estimating σ .

5.12.2

Problems with Standardized Effect Sizes

Many researchers use Cohen's standardized effect sizes in planning a study. This has the advantage of not requiring pilot data. But such effect sizes are not biologically meaningful and may hide important issues⁶⁰. Studies should be designed on the basis of effects that are relevant to the investigator and human subjects. If, for example, one plans a study to detect a one standard deviation (SD) difference in the means and the SD is large, one can easily miss a biologically important difference that happened to be much less than one SD in magnitude. Note that the SD is a measure of how subjects disagree with one another, not a measure of an effect (e.g., the shift in the mean). One way to see that standardized effect sizes are problematic is to note that if one were to make the measurements more noisy, the SD will increase and the purported clinically important difference to detect will increase proportionately.

5.12.3

Choice of Effect Size

If a study is designed to detect a certain effect size with a given power, the effect size should never be the observed effect from another study, which may be estimated with error and be overly optimistic. The effect size to use in planning should be the clinically or biologically relevant effect one would regret missing. Usually the only information from prior studies that is useful in sample size estimation are (in the case of a continuous response variable with a symmetric distribution) estimates of the standard deviation or the correlation between two measurements on the same subject measured at two different times, or (in the case of a binary or time to event outcome) event probabilities in control subjects. An excellent resource by Senn for understanding effect sizes in power calculations may be found at bit.ly/ssenn-effect.

- Effect size for power/sample size calculation is **never** an observed effect in previous data
- It is not the difference we believe is true
- It is the difference you would not like to miss
- Clinically relevant to patients or at least to physiology
- Not *greater than clinically relevant*

5.12.4

Multiple Estimands and Hypotheses

In many experiments there are more than one estimand (what is to be estimated based on the question of interest) or hypothesis. Some frequentist statisticians and biomedical investigators believe that in such situations the familywise error probability should be controlled^x. This probability is the probability of rejecting *any* null hypothesis given that *all* null hypotheses are true. One may accomplish this by testing every hypothesis at the α^* level, where the constant $\alpha^* < \alpha$ is chosen so that the overall type one error is α , or one may elect to differentially “spend α ” by, for example, setting $\alpha = 0.04$ for

^xAs stated elsewhere, multiplicity adjustments are a byproduct of faults in frequentist inference and are completely arbitrary.

a primary hypothesis and $\alpha = 0.01$ for a less important secondary analysis. Another alternative is closed testing procedures whereby later hypotheses can be tested at less stringent α levels as long as all earlier hypotheses were rejected. Unfortunately there is no unique path to deriving multiplicity adjustments, and they have the odd property of requiring one to be more stringent in assessing evidence for one hypothesis just because one had other hypotheses.

An alternative, and what we believe to be more reasonable, view is by Cook and Farewell²² who stated that if a study has more than one question and each question is to be answered on its own, there is no need for a multiplicity adjustment. This is especially true if a strong priority ordering for hypotheses is stated in advance. For example, an investigator may specify three hypotheses about efficacy of a treatment for the following endpoints in a cardiovascular trial, sorted from most important to least important: overall mortality, cardiovascular mortality, and cardiovascular death or myocardial infarction. As long as the researcher always reports all of the results in context, in this pre-specified order, each P -value can stand on its own.

Contrast this with an exploratory study in which the hypothesis is essentially that there exists an endpoint for which the treatment is effective. One should expect to have to employ a conservative multiplicity adjustment in that situation, e.g., Bonferroni's inequality.

Consider a frequentist study with four efficacy endpoints and corresponding P -values in the given pre-specified priority order: all-cause mortality ($P = 0.09$), stroke ($P = 0.01$), myocardial infarction ($P = 0.06$), hospitalization ($P = 0.11$)

- OK to quantify evidence against each of the 4 null hypotheses if all 4 reported in context, using the pre-specified order with separate interpretations
- Reasonable conclusion: With the current sample size, there is little evidence to reject the supposition that treatment does not lower mortality. There is evidence against the supposition that treatment does not lower the rate of stroke. . . .
- Contrast with a report that merely reports the stroke result, essentially saying “there exists an endpoint for which treatment is effective”
- Example Bayesian statement: Treatment probably (0.92) lowers mortality, probably (0.995) lowers the rate of stroke, probably (0.96) lowers MI, and probably (0.96) lowers hospitalization (posterior probabilities of a treatment benefit are in

parentheses)^y.

- Perhaps better: create a 5 or more level ordinal endpoint and use the worst event as the response variable
 - $Y = 0, 1, 2, 3, 4$ corresponding to no event, hospitalization, MI, stroke, death
 - Interpretation 1: treatment lowered the odds of an outcome or a worse outcome by a factor of 0.8
 - Interpretation 2: chance of MI, stroke, or death with treatment estimated as 0.167 and for control as 0.2
chances of stroke or death: 0.082, 0.1
 - Bayesian probability of treatment benefit = $P(\text{OR} < 1) = 0.998$
- See Section 3.6 for how properties of ordinal scales relate to power

5.12.5

Study Design Big Picture

- Choose the right question or estimand
- Think hard about subject/animal selection criteria
- Decide whether you are doing a pilot study or a more definitive study
 - Pilot study is not used to nail down the effect of an intervention
 - Is to show you can make the measurements, carry out the intervention, refine existing measurements, enroll enough subjects per month, etc.
 - Power is not relevant
 - May size the pilot study to be able to estimate something simple with precision (proportion, SD); adequate estimation of SD for continuous Y is important for sample size calculation for the full study

^yA Bayesian analysis would quickly add the (lower) probabilities that treatment lowers event rates by more than a trivial amount. Bayesian methods can also compute the probability that any 2 of the efficacy targets were achieved, and the expected number of targets hit—in this case $0.92+0.995+0.96+0.96=3.8$ of 4 targets hit.

- Make sure the design can answer that question were the sample size 1,000,000 subjects
- Decide whether you are required to have a fixed sample size
 - If budgeting is flexible, use fully sequential design and stop when evidence is adequate^z
 - For fixed budget/study duration be realistic about the effect size
- If there is inadequate budget for detecting the minimal clinically important effect with high probability, be realistic about stating the study's likely yield
 - Best: compute the likely margin of error for the primary estimand
 - Next best: compute the power that will be achieved with the limited sample size
- Choose a response variable Y that answers the question and has the greatest frequentist or Bayesian power (section 3.6))
 - Example: primary interest is patient's inability to function physically on a 0-100 scale (100=bedridden)
 - Some patients will be too sick to have their function assessed and some will die
 - Define $Y=0-100$ overridden with 101 for too sick or 102 for death^a
 - Analyze with the proportional odds model
 - Interpretation:
 - * Primary endpoint is degree of functional disability, penalized by death or being physically unable to be assessed
 - * Proportional odds model provides an overall odds ratio for treatments B:A (ratio of odds that $Y \geq j$ for any j)
 - * Model can also be used to estimate the median disability, where sickness or death will shift the median to the right a little

^zBayesian sequential designs require no penalty for infinitely many such data looks. See fharrell.com/post/bayes-seq.

^aOrdinal analysis will not be affected by placing the clinical events at 1001 and 1002 or any other levels that are higher than the highest functional disability level.

- * May also be summarized by estimating for each treatment the probability that a patient has level 50 functional disability or worse where “or worse” means 51-100, too sick, or dead, after estimating the overall odds ratio for treatment
- Use multiple measurements over time to increase power/precision and to allow more questions to be answered
- Greatest power comes from having a continuous Y or ordinal Y with many well-populated levels, where Y is also measured at baseline and is adjusted for as a covariate in ANCOVA, allowing for a smooth nonlinear effect (without assuming the slope is 1.0 as is assumed by change-from-baseline analysis)
- Never use change from baseline as the response variable except in a non-randomized pre-post design (the weakest of all designs)
- If treatment is short-term and wears off, fully using each subject as her own control in a randomized crossover study may be ideal
- For a parallel-group randomized study, accurately collect key baseline variables that explain outcome heterogeneity
- For an observational study, accurately capture a host of baseline variables that are likely to result in adequate confounder adjustment
 - don’t merely rationalize that variables available in an existing dataset are adequate
- Use a research data management tool such as REDCap that allows for extensive data quality checking
- Don’t forget the many subject selection, ethical, and good clinical practice issues
- Recommended reading: Hulley et al. *Designing Clinical Research*⁴⁸

5.13

One-Sample t -Test Revisited

5.13.1

Study Description

- Compare the effects of two soporific drugs (optical isomers of hyoscyamine hydrobromide)
- Crossover study
- Each subject receives placebo run-in, then Drug 1, then Drug 2
- Investigator may not have randomized order of treatments
- Dependent variable: Number of hours of increased sleep when compared to a placebo run-in period (raw data not shown)
- Drug 1 given to n subjects, Drug 2 given to same n subjects
- Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
 - $H_0 : \mu_d = 0$ where $\mu_d = \mu_1 - \mu_2$
 - $H_1 : \mu_d \neq 0$

5.13.2

Power and Sample Size

- Pilot study or previous published research shows the standard deviation of the difference (σ_d) is 1.2 hours
- Determine the number of subjects needed for several value of effect size Δ ($\Delta = |\mu_1 - \mu_2|$) with 0.9 power, $\alpha = 0.05$

Δ	0.5	1	1.5	2
n	62	16	8	5

- If Drug 1 (or 2) increases sleep by 1.5 hours more than Drug 2 (or 1), by enrolling 8 subjects we will have 0.9 power to detect an association.
- More powerful than the two sample test (need 10 subjects in each group for $\Delta = 3.0$ hours)

5.13.3

Collected Data

Here are the data for the 10 subjects. This is the R built-in dataset `sleep`.

Subject	Drug 1	Drug 2	Diff (2-1)
1	0.7	1.9	1.2
2	-1.6	0.8	2.4
3	-0.2	1.1	1.3
4	-1.2	0.1	1.3
5	-0.1	-0.1	0.0
6	3.4	4.4	1.0
7	3.7	5.5	1.8
8	0.8	1.6	0.8
9	0.0	4.6	4.6
10	2.0	3.4	1.4
Mean	0.75	2.33	1.58
SD	1.79	2.0	1.2

```

drug1 <- c(.7, -1.6, -.2, -1.2, -.1, 3.4, 3.7, .8, 0, 2)
drug2 <- c(1.9, .8, 1.1, .1, -.1, 4.4, 5.5, 1.6, 4.6, 3.4)
d <- data.frame(Drug=c(rep('Drug 1', 10), rep('Drug 2', 10),
                      rep('Difference', 10)),
                  extra=c(drug1, drug2, drug2 - drug1))
w <- data.frame(drug1, drug2, diff=drug2 - drug1)

ggplot(d, aes(x=Drug, y=extra)) +    # Fig. 5.11
  geom_boxplot(col='lightyellow1', alpha=.3, width=.5) +
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
  stat_summary(fun.y=mean, geom="point", col='red', shape=18, size=5) +
  geom_segment(data=w, aes(x='Drug 1', xend='Drug 2', y=drug1, yend=drug2),
               col=gray(.8)) +
  
```

```
geom_segment(data=w, aes(x='Drug 1', xend='Difference', y=drug1, yend=drug2 -
    drug1),
             col=gray(.8)) +
  xlab('') + ylab('Extra Hours of Sleep') + coord_flip()
```

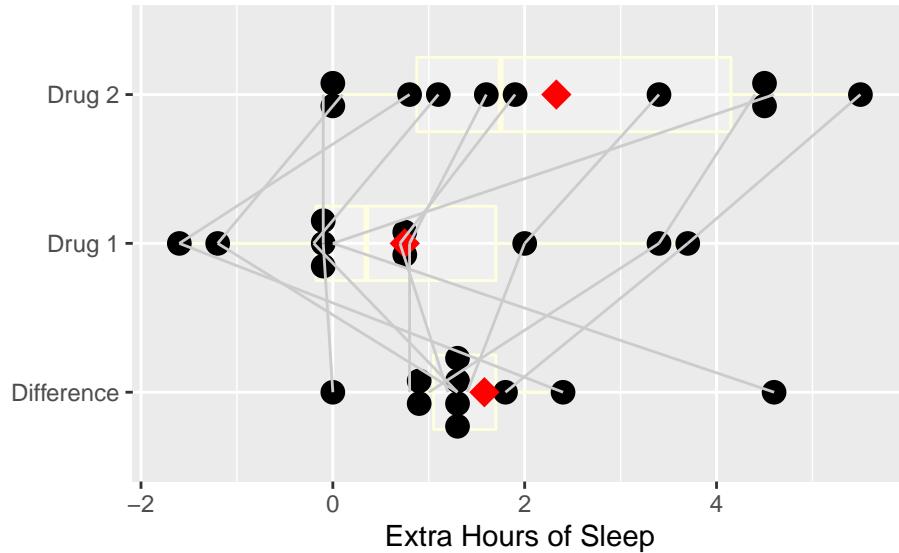


Figure 5.11: Raw data and box plots for paired data and their paired differences, with lines connecting points from the same subject. Diamonds depict means.

5.13.4

Statistical Test

```
with(d, t.test(drug1, drug2, paired=TRUE))
```

```
Paired t-test

data: drug1 and drug2
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

- Interpretation

- A person who takes Drug 2 sleeps on average 1.58 hours longer (0.95 CI: [0.70, 2.46]) than a person who takes Drug 1

5.14

Comprehensive Example: Crossover Design and Analysis

- In the previous example, it was not clear if the order of placebo, Drug 1, and Drug 2 was the same for every patient
- In a crossover design, each patient receives both drugs
 - Can serve as own control
 - Effectively adjusts for all baseline variables without measuring them^b
 - Order is randomized
- Carryover effects
 - Def: An effects that carries over from one experimental condition to another
 - Need a washout period between drugs to remove carryover effects
 - Time to remove carryover effects should be based on biology, not statistics
 - Statistical tests for carryover effects are often not precise enough to make definitive conclusions (see example)
 - The test for carryover is correlated with the overall test of efficacy
 - Pre-testing for carryover then deciding whether to only use phase 1 data results in a huge inflation of type I error in the test for efficacy

5.14.1

Study Description

- Compare the effects of two soporific drugs.

^bIf there is no interaction between covariate and treatment order

- Each subject either (1) starts with Drug 1 and crosses over to Drug 2 or (2) starts with Drug 2 and crosses over to Drug 1
 - No placebo run-in in this example
 - Order randomly assigned
 - Suitable period of time (~ 5 half-lives) between drug crossovers to washout effects of previous drug
- Dependent variable: Number of hours of sleep on each drug
- Drug 1 given to n subjects, Drug 2 given to same n subjects
- Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
 - $H_0 : \mu_d = 0$ where $\mu_d = \mu_1 - \mu_2$
 - $H_1 : \mu_d \neq 0$

5.14.2

Power and Sample Size

- Pilot study or previous published research shows the standard deviation of the difference (σ_d) is 1.2 hours
- Determine the number of subjects needed for several value of effect size Δ ($\Delta = |\mu_1 - \mu_2|$) with 0.9 power, $\alpha = 0.05$
- Assume no carryover effects

Δ	0.5	1	1.5	2
n	62	16	8	5

- If Drug 1 (or 2) increases sleep by 1.5 hours more than Drug 2 (or 1), by enrolling 8 subjects we will have 0.9 power to detect an association.
- Same power calculation as paired t -test

5.14.3**Collected Data**

Subject	Drug 1	Drug 2	Diff (2-1)
1	8.7	9.9	1.2
2	6.4	8.8	2.4
3	7.8	9.1	1.3
4	6.8	8.1	1.3
5	7.9	7.9	0.0
6	11.4	12.4	1.0
7	11.7	13.5	1.8
8	8.8	9.6	0.8
9	8.0	12.6	4.6
10	10.0	11.4	1.4
Mean	8.75	10.33	1.58
SD	1.79	2.0	1.2

5.14.4**Statistical Tests**

```
drug1 ← c(87, 64, 78, 68, 79, 114, 117, 88, 80, 100)/10
drug2 ← c(99, 88, 91, 81, 79, 124, 135, 96, 126, 114)/10
t.test(drug1, drug2, paired=TRUE)
```

```
Paired t-test

data: drug1 and drug2
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

- Interpretation

- A person who takes Drug 2 sleeps on average 1.58 hours longer (0.95 CI: [0.70, 2.50]) than a person who takes Drug 1

5.14.5

Carryover Effects

- Is there any evidence for a carryover effect?
- Assume that the first 5 subjects received Drug 1 first and the second 5 subjects received drug 2 first
- If we assume there are no carryover effects, then the mean difference in sleep for subjects receiving drug 1 first should be *the same* as the mean difference for subjects receiving drug 2 first
- Assessing carryover effect distorts the efficacy analysis inference
- Null hypothesis is that there are no carryover effects
- Can rearrange the difference data to clarify the structure

Subject	Drug 1 First	Drug 2 First
1		1.2
2		2.4
3		1.3
4		1.3
5		0.0
6		1.0
7		1.8
8		0.8
9		4.6
10		1.4
Mean	1.24	1.92
SD	0.85	1.55

For this design we might expect the variance of the differences to be the same for both orders, so we use the equal-variance t -test.

```
# Unpaired t-test
t.test((drug2 - drug1)[1:5], (drug2 - drug1)[6:10], var.equal=TRUE)
```

Two Sample t-test

```
data: (drug2 - drug1)[1:5] and (drug2 - drug1)[6:10]
t = -0.86152, df = 8, p-value = 0.414
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.500137 1.140137
sample estimates:
mean of x mean of y
1.24      1.92
```

- Interpretation
 - Large P -value has no interpretation
 - With 0.95 confidence, the carryover effect is between [-2.5 and 1.1] hours, which is not scientifically convincing either way
 - In general, be very cautious when the null hypothesis is something you want to fail to reject in order to validate your analysis method
 - * Tests of normality are sometimes used to validate using a parametric over a non-parametric test
 - * There are also statistical tests for equal variance
 - * Both tests may be unreliable and will distort the final inference that is conditional on preassessments being correct
- As [Stephen Senn](#) has warned, be wary of doing anything about empirically quantified carryover effects, as the carryover effect estimate has a correlation of $\frac{1}{2}$ with the overall treatment effect estimate, causing the carryover test to ruin the operating characteristics of the treatment test

5.14.6

Bayesian Analysis

- Reasonable to put prior knowledge on parameters, especially carryover event
- Reasonable to restrict carryover effect to be less than the treatment effect

- For related discussions and references for Bayesian crossover analysis see
bit.ly/datamethods-bbr7

Chapter 6

Comparing Two Proportions



6.1

Overview

- Compare dichotomous independent variable with a dichotomous outcome
 - Independent variables: Exposed/Not, Treatment/Control, Knockout/Wild Type, etc.
 - Outcome (dependent) variables: Diseased/Not or any Yes/No outcome
- Continuous outcomes often dichotomized for analysis (bad idea)
 - Consider *t*-tests (Chapter 5) or Non-parameteric methods (Chapter 7)

6.2

Normal-Approximation Test

- Two independent samples

	Sample 1	Sample 2
Sample size	n_1	n_2
Population probability of event	p_1	p_2
Sample probability of event	\hat{p}_1	\hat{p}_2

- Null Hypothesis, $H_0 : p_1 = p_2 = p$

- Estimating the variance
 - Variance of $\hat{p}_i = p_i(1 - p_i)/n_i$ for $i = 1, 2$
 - Variance of $(\hat{p}_1 - \hat{p}_2)$ is the sum of the variances, which under H_0 is

$$p(1 - p)[\frac{1}{n_1} + \frac{1}{n_2}]$$

- We estimate this variance by plugging \hat{p} into p , where

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the pooled estimate of the probability under $H_0 : p_1 = p_2 = p$

- Test statistic which has approximately a normal distribution under H_0 if $n_i\hat{p}_i$ are each large enough:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})[\frac{1}{n_1} + \frac{1}{n_2}]}}$$

- To test H_0 we see how likely it is to obtain a z value as far or farther out in the tails of the normal distribution than z is
- We don't recommend using the continuity correction

- Example:

Test whether the population of women whose age at first birth ≤ 29 has the same probability of breast cancer as women whose age at first birth was ≥ 30 . This

dichotomization is highly arbitrary and we should really be testing for an association between age and cancer incidence, treating age as a continuous variable.

- Case-control study (independent and dependent variables interchanged); p_1 = probability of age at first birth ≥ 30 , etc.

	<u>with Cancer</u>	<u>without Cancer</u>
Total # of subjects	3220(n_1)	10245(n_2)
# age ≥ 30	683	1498

Sample probabilities $0.212(\hat{p}_1)$ $0.146(\hat{p}_2)$

Pooled probability $\frac{683+1498}{3220+10245} = 0.162$

- Estimate the variance

$$-\text{variance}(\hat{p}_1 - \hat{p}_2) = \hat{p}(1 - \hat{p}) \times \left[\frac{1}{n_1} + \frac{1}{n_2} \right] = 5.54 \times 10^{-5}$$

$$-SE = \sqrt{\text{variance}} = 0.00744$$

- Test statistic

$$-z = \frac{0.212 - 0.146}{0.00744} = 8.85$$

- 2-tailed P -value is $< 10^{-4}$

```
n1 <- 3220;      n2 <- 10245
p1 <- 683 / n1; p2 <- 1498 / n2
pp <- (n1 * p1 + n2 * p2) / (n1 + n2)
se <- sqrt(pp * (1 - pp) * (1 / n1 + 1 / n2))
z <- (p1 - p2) / se
pval <- 2 * (1 - pnorm(abs(z)))
round(c(p1=p1, p2=p2, pooled=pp, se=se, z=z, pval=pval), 4)
```

p1	p2	pooled	se	z	pval
0.2121	0.1462	0.1620	0.0074	8.8527	0.0000

- We do not use a t -distribution because there is no σ to estimate (and hence no “denominator d.f.” to subtract)

6.3

 χ^2 Test

- If z has a normal distribution, z^2 has a χ^2 distribution with 1 d.f. (are testing a single difference against zero)
- The data we just tested can be shown as a 2×2 contingency table

	Cancer +	Cancer -	
Age ≤ 29	2537	8747	11284
Age ≥ 30	683	1498	2181
	3220	10245	13465

- In general, the χ^2 test statistic is given by

$$\sum_{ij} \frac{(\text{Obs}_{ij} - \text{Exp}_{ij})^2}{\text{Exp}_{ij}}$$

- Obs_{ij} is the observed cell frequency for row i column j
- Exp_{ij} is the expected cell frequency for row i column j
 - Expected cell frequencies calculating assuming H_0 is true
 - $\text{Exp}_{ij} = \frac{\text{row } i \text{ total} \times \text{column } j \text{ total}}{\text{grand total}}$
 - e.g. $\text{Exp}_{11} = \frac{11284 \times 3220}{13465} = 2698.4$

- For 2×2 tables, if the observed cell frequencies are labeled $\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array}$ the χ^2 test statistic simplifies to

$$\frac{N[ad - bc]^2}{(a+c)(b+d)(a+b)(c+d)},$$

where $N = a + b + c + d$. Here we get $\chi^2_1 = 78.37$

- 78.37 is z^2 from above!

```
x <- matrix(c(2537, 8747, 683, 1498), nrow=2, byrow=TRUE)
x
```

```
[,1] [,2]
[1,] 2537 8747
[2,] 683 1498
```

```
chisq.test(x, correct=FALSE)
```

```
Pearson's Chi-squared test

data: x
X-squared = 78.37, df = 1, p-value < 2.2e-16
```

```
# Also compute more accurate P-value based on 1M Monte-Carlo simulations
chisq.test(x, correct=FALSE, simulate.p.value=TRUE, B=1e6)
```

```
Pearson's Chi-squared test with simulated p-value (based on 1e+06
replicates)

data: x
X-squared = 78.37, df = NA, p-value = 1e-06
```

- Don't need Yates' continuity correction
- Note that even though we are doing a 2-tailed test we use only the right tail of the χ^2_1 distribution; that's because we have squared the difference when computing the statistic, so the sign is lost.
- This is the ordinary Pearson χ^2 test

6.4

Fisher's Exact Test

- Is a misnomer in the sense that it computes probabilities exactly, with no normal approximation, but only after changing what is being tested to condition on the number of events and non-events
- Because frequencies are discrete and because of the conditioning, the test is conservative (P -values too large)
- Is exact only in the sense that actual type I error probability will not **exceed** the nominal level
- The ordinary Pearson χ^2 works fine (even when an expected cell frequency is as low as 1.0, contrary to popular belief)
- We don't use Yates' continuity correction because it was developed to make the normal approximation test yield P -values that are more similar to Fisher's test, i.e., to be more conservative
- The attempt to obtain exact unconditional P -values for the simple 2×2 contingency table has stumped frequentist statisticians for many decades¹⁷
- By contrast, Bayesian posterior probabilities for the true unconditional quantity of interest are exact
 - Frequentist confidence limits and P -values are approximate because they use the sample space, and the sample space is discrete when the response variable is categorical
 - Bayes does not consider the sample space, only the parameter space, which is almost always continuous
- See stats.stackexchange.com/questions/14226 for discussion

6.5

Sample Size and Power for Comparing Two Independent Samples

- Power \uparrow as
 - $n_1, n_2 \uparrow$
 - $\frac{n_2}{n_1} \rightarrow 1.0$ (usually)
 - $\Delta = |p_1 - p_2| \uparrow$
 - $\alpha \uparrow$
- There are approximate formulas such as the recommended methods in Altman based on transforming \hat{p} to make it have a variance that is almost independent of p
- Example:

Using current therapy, 0.5 of the population is free of infection at 24 hours. Adding a new drug to the standard of care is expected to increase the percentage infection-free to 0.7. If we randomly sample 100 subjects to receive standard care and 100 subjects to receive the new therapy, what is the probability that we will be able to detect a certain difference between the two therapies at the end of the study?

$$p_1 = .5, p_2 = .7, n_1 = n_2 = 100$$

results in a power of 0.83 when $\alpha = 0.05$

```
require(Hmisc)
bpower(.5, .7, n1=100, n2=100)
```

Power 0.8281098

- When computing sample size to achieve a given power, the sample size \downarrow when
 - power \downarrow

$$-\frac{n_2}{n_1} \rightarrow 1.0$$

- $\Delta \uparrow$

- $\alpha \uparrow$

- Required sample size is a function of both p_1 and p_2

- Example:

How many subjects are needed to detect a 0.8 fold decrease in the probability of colorectal cancer if the baseline probability of cancer is 0.0015? Use a power of 0.8 and a type-I error probability of 0.05.

$$p_1 = 0.0015, p_2 = 0.8 \times p_1 = 0.0012, \alpha = 0.05, \beta = 0.2$$

$$n_1 = n_2 = 235, 147$$

(Rosner estimated 234,881)

```
bsamsiz(.0015, 0.8 * .0015, alpha=0.05, power=0.8)
```

n1	n2
235147.3	235147.3

Formulas for power and sample size may be seen as R code found at
github.com/harrelfe/Hmisc/blob/master/R/bpower.s.

6.6

Confidence Interval

An approximate $1 - \alpha$ 2-sided CL is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

where $z_{1-\alpha/2}$ is the critical value from the normal distribution (1.96 when $\alpha = 0.05$).

The CL for the number of patients needed to be treated to save one event may simply be obtained by taking the reciprocal of the two confidence limits.^a

^aIf a negative risk reduction is included in the confidence interval, set the NNT to ∞ for that limit instead of quoting a negative NNT. There is more to this; see bit.ly/datamethods-nnt.

6.7

Sample Size for a Given Precision

- Goal: Plan a study so that the margin of error is sufficiently small
- The margin of error (δ) is defined to be half of the confidence interval width. For two proportions,

$$\delta = z_{1-\alpha/2} \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Basing the sample size calculations on the margin of error can lead to a study that gives *scientifically* relevant results even if the results are not *statistically* significant.
- Example: Suppose that the infection risk in a population is 0.5 and a reduction to 0.4 is believed to be a large enough reduction that it would lead to a change in procedures. A study of a new treatment is planned so that enough subjects will be enrolled for the margin of error is 0.05. Consider these two possible outcomes:
 1. The new treatment is observed to decrease infections by 0.06 (0.95 CI: [0.11, 0.01]). The confidence interval does not contain 0, so we have indirect evidence^b that the new treatment is effective at reducing infections. 0.1 is also within the confidence interval limits.
 2. The new treatment is observed to decrease infections by only 0.04 (0.95 CI: [0.09, -0.01]). The confidence interval now contains 0, so we do not have enough evidence to reject the supposition that there is no effect of the treatment on reducing infections if we are bound to an arbitrary $\alpha = 0.05$. However, the confidence interval also does not contain 0.10, so we are able to indirectly rule out a *scientifically* relevant decrease in infections.

- For fixed $n_1 = n_2 = n$, confidence intervals for proportions have the maximum width, when $p_1 = p_2 = 0.5$. This can be shown by:
 - Recall that the variance formula for the difference in two proportions when calculating a confidence interval is

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

^bTo obtain direct evidence requires Bayesian posterior probabilities.

- When $p_1 = p_2 = p$ and $n_1 = n_2 = n$, the variance formula simplifies to

$$\frac{p(1-p)}{n} + \frac{p(1-p)}{n} = 2\frac{p(1-p)}{n}$$

- Then, for any fixed value of n (e.g. $n = 1$ or 10), $2\frac{p(1-p)}{n}$ is largest when $p = 0.5$. With $p = 0.5$, the variance formula further simplifies to

$$2\frac{.25}{n} = \frac{1}{2n}$$

- Using $\alpha = 0.05$ ($z_{1-\alpha/2} = 1.96$), the worst-case margin of error will be

$$\delta = 1.96\sqrt{\frac{1}{2n}}$$

- By solving for n , we can rearrange this formula to be

$$n = \frac{1.92}{\delta^2}$$

- This formula then gives the number of subjects needed in each group n to obtain a given margin of error δ . For a margin of error of 0.05 ($\delta = 0.05$), $n = \frac{1.92}{0.05^2} = 768$ subjects in each group.

```
diff <- .05
qnorm(.975)^2 / 2 / (diff ^ 2)
```

```
[1] 768.2918
```

6.8

Relative Effect Measures

- We have been dealing with risk differences which are measures of absolute effect
- Measures of relative effect include risk ratios and odds ratios
- Risk ratios are easier to interpret but only are useful over a limited range of prognosis (i.e., a risk factor that doubles your risk of lung cancer cannot apply to a subject having a risk above 0.5 without the risk factor)
- Odds ratios can apply to any subject
- In large clinical trials treatment effects on lowering probability of an event are often constant on the odds ratio scale
- $\text{OR} = \text{Odds ratio} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$
- Testing $H_0: \text{OR}=1$ is equivalent to testing $H_0 : p_1 = p_2$
- There are formulas for computing confidence intervals for odds ratios
- Odds ratios are most variable when one or both of the probabilities are near 0 or 1
- We compute CLs for ORs by anti-logging CLs for the log OR
- In the case where $p_1 = p_2 = 0.05$ and $n_1 = n_2 = n$, the standard error of the log odds ratio is approximately $\sqrt{\frac{42.1}{n}}$
- The common sample size n needed to estimate the true OR to within a factor of 1.5 is 984 with ps in this range
- To show the multiplicative margins of error^c for a range of sample sizes and values of p . For each scenario, the margin of error assumes that both unknown probability estimates equal p .

^cValue by which to multiply the observed odds ratio to obtain the upper 0.95 confidence limit or to divide the observed odds ratio to obtain the lower 0.95 limit

```

require(ggplot2)
d <- expand.grid(n=c(seq(10, 1000, by=10), seq(1100, 50000, by=100)),
                  p=c(.02, .05, .075, .1, .15, .2, .25, .3, .4, .5))
d$selor <- with(d, sqrt(2 / (p * (1 - p) * n)))
d$mmoe <- with(d, exp(qnorm(0.975) * selor))
mb <- c(1, 1.25, 1.5, 2, 2.5, 3, 4, 5, 10, 20, 30, 40, 50, 100, 400)
ggplot(aes(x=n, y=mmoe, color=factor(p)), data=d) + # Fig. 6.1
  geom_line() +
  scale_x_log10(breaks=c(10,20,30,50,100,200,500,1000,2000,5000,10000,
                         20000,50000)) +
  scale_y_log10(breaks=mb, labels=as.character(mb)) +
  xlab(expression(n)) + ylab('Multiplicative Margin of Error for OR') +
  guides(color=guide_legend(title=expression(p)))

```

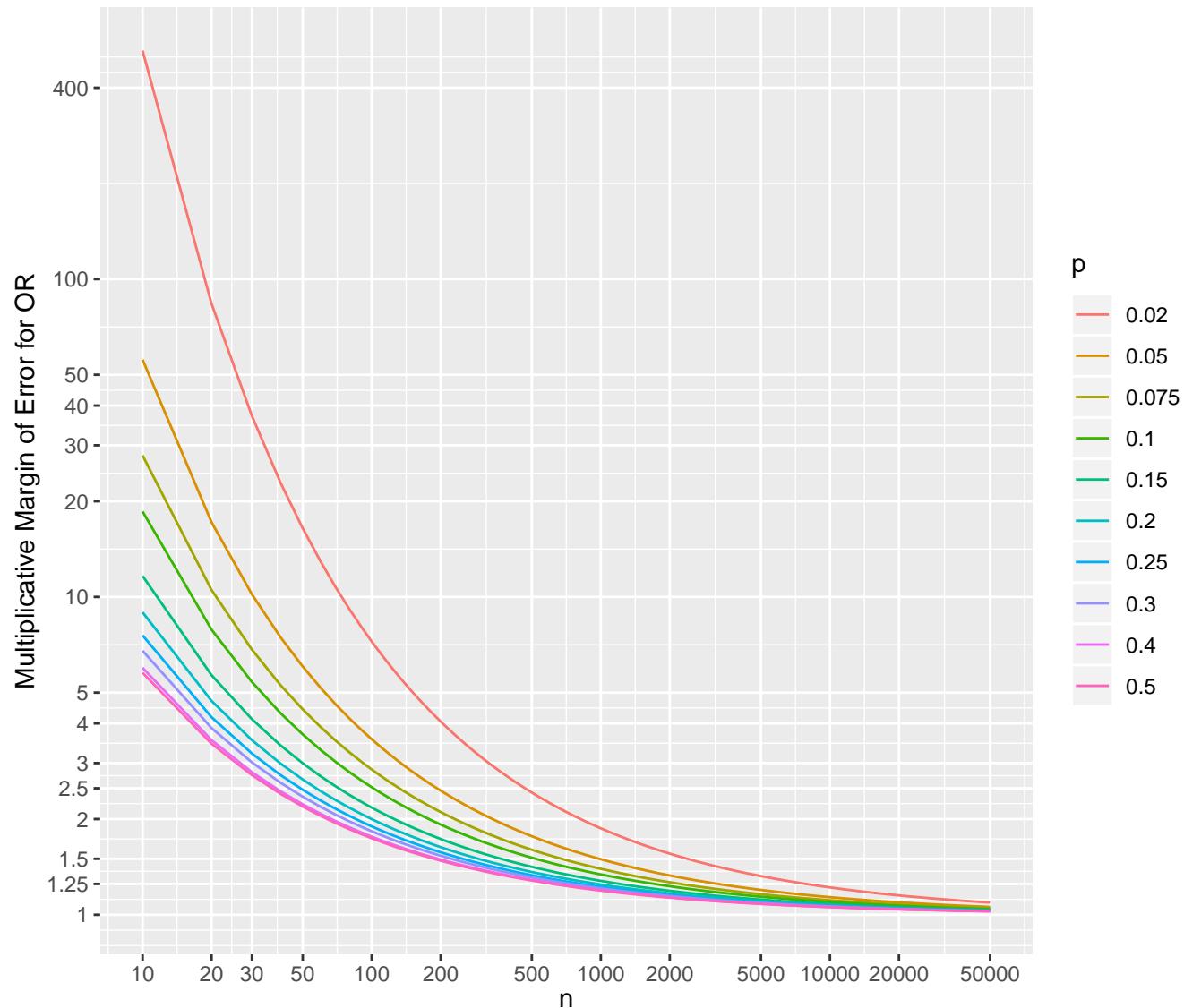


Figure 6.1: Multiplicative margin of error related to 0.95 confidence limits of an odds ratio, for varying n and p (different curves), assuming the unknown true probability in each group is no lower than p

6.9

Comprehensive example

6.9.1

Study Description

- Consider patients who will undergo coronary artery bypass graft surgery (CABG)
- Mortality risk associated with open heart surgery
- Study question: Do emergency cases have a surgical mortality that is different from that of non-emergency cases?
- Population probabilities
 - p_1 : Probability of death in patients with emergency priority
 - p_2 : Probability of death in patients with non-emergency priority
- Statistical hypotheses
 - $H_0 : p_1 = p_2$ (or OR = 1)
 - $H_1 : p_1 \neq p_2$ (or OR $\neq 1$)

6.9.2

Power and Sample Size

- Prior research shows that just over 0.1 of surgeries end in death
- Researchers want to be able to detect a 3 fold increase in risk
- For every 1 emergency priority, expect to see 10 non-emergency
- $p_1 = 0.3$, $p_2 = 0.1$, $\alpha = 0.05$, and power = 0.90

- Calculate sample sizes using the PS software for these values and other combinations of p_1 and p_2

(p_1, p_2)	(0.3, 0.1)	(0.4, 0.2)	(0.03, 0.01)	(0.7, 0.9)
n_1	40	56	589	40
n_2	400	560	5890	400

Check PS calculations against the R `Hmisc` package's `bsamsize` function.

```
round(bsamsize(.3, .1, fraction=1/11, power=.9))
```

n1	n2
40	399

```
round(bsamsize(.4, .2, fraction=1/11, power=.9))
```

n1	n2
56	561

```
round(bsamsize(.7, .9, fraction=1/11, power=.9))
```

n1	n2
40	399

6.9.3

Collected Data

In-hospital mortality figures for emergency surgery and other surgery

Surgical Priority	Discharge Status	
	Dead	Alive
Emergency	6	19
Other	11	100

- $\hat{p}_1 = \frac{6}{25} = 0.24$

- $\hat{p}_2 = \frac{11}{111} = 0.10$

6.9.4

Statistical Test

```
n1 ← 25;      n2 ← 111
p1 ← 6 / n1; p2 ← 11 / n2
or ← p1 / (1 - p1) / (p2 / (1 - p2))
or
```

```
[1] 2.870813
```

```
# Standard error of log odds ratio:
selor ← sqrt(1 / (n1 * p1 * (1 - p1)) + 1 / (n2 * p2 * (1 - p2)))
# Get 0.95 confidence limits
cls ← exp(log(or) + c(-1, 1) * qnorm(0.975) * selor)
cls
```

```
[1] 0.946971 8.703085
```

```
tcls ← paste0(round(or, 2), ', (0.95 CI: [', round(cls[1], 2),
              ', ', round(cls[2], 2), '])')
# Multiplying a constant by the vector -1, 1 does +/- 
x ← matrix(c(6, 19, 11, 100), nrow=2, byrow=TRUE)
x
```

```
[,1] [,2]
[1,]    6   19
[2,]   11  100
```

```
chisq.test(x, correct=FALSE)
```

```
Pearson's Chi-squared test

data: x
X-squared = 3.7037, df = 1, p-value = 0.05429
```

- Interpretation

- Compare odds of death in the emergency group $\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right)$ to odds of death in non-emergency group $\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right)$
- Emergency cases have 2.87 (0.95 CI: [0.95, 8.7]) fold increased odds of death during surgery compared to non-emergency cases.

Fisher's Exact Test

Observed marginal totals from emergency surgery dataset

	Dead	Alive	
Emergency	a	b	25
Other	c	d	111
	17	119	136

- With fixed marginal totals, there are 18 possible tables ($a = 0, 1, \dots, 17$)
- Can calculate probability of each of these tables
 - p -value: Probability of observing data as extreme or more extreme than we collected in this experiment
- Exact test: p -value can be calculated “exactly” (not using the χ^2 distribution to approximate the p -value)

```
fisher.test(x)
```

```
Fisher's Exact Test for Count Data

data: x
p-value = 0.08706
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.7674155 9.6831351
sample estimates:
odds ratio
2.843047
```

Note that the odds ratio from Fisher's test is a conditional maximum likelihood estimate, which differs from the unconditional maximum likelihood estimate we obtained earlier.

- Fisher's test more conservative than Pearson's χ^2 test (larger P -value)

6.10

Logistic Regression for Comparing Proportions

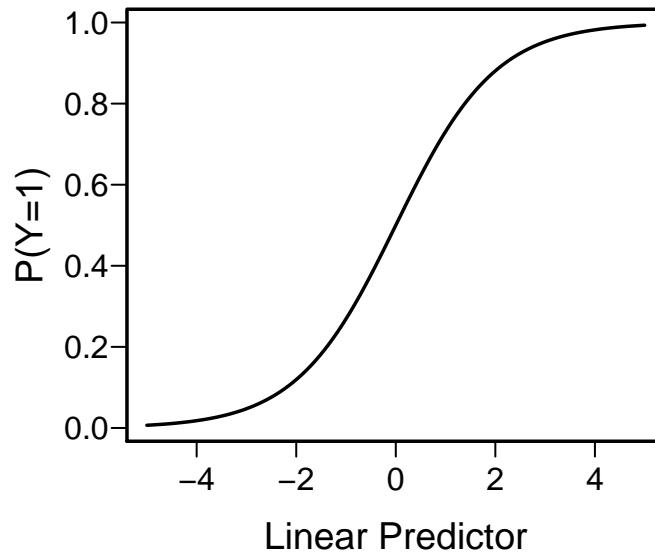


- When comparing ≥ 2 groups on the probability that a categorical outcome variable will have a certain value observed (e.g., $P(Y = 1)$), one can use the Pearson χ^2 test for a contingency table (or the less powerful Fisher's "exact" test)
- Such analyses can also be done with a variety of regression models
- It is necessary to use a regression model when one desires to analyze more than the grouping variable, e.g.
 - analyze effects of two grouping variables
 - adjust for covariates
- For full generality, the regression model needs to have no restrictions on the regression coefficients
 - Probabilities are restricted to be in the interval $[0, 1]$ so an additive risk model cannot fit over a broad risk range
 - Odds ($\frac{p}{1-p}$) are restricted to be in $[0, \infty]$
 - Log odds have no restrictions since they can be in $[-\infty, \infty]$
- So log odds is a good basis for regression analysis of categorical Y
 - default assumption of additivity of effects needs no restrictions
 - will still translate to probabilities in $[0, 1]$
- The binary logistic regression model is a model to estimate the probability of an event as a flexible function of covariates
- Let the outcome variable Y have the values $Y = 0$ (non-event) or $Y = 1$ (event)

$$\text{Prob}(Y = 1|X) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots))} \quad (6.1)$$

- The sum inside the inner () is called the linear predictor (LP)
- The binary logistic model relates LP (with no restrictions) to the event probability as so:

```
lp <- seq(-5, 5, length=150)
plot(lp, plogis(lp), xlab='Linear Predictor', ylab='P(Y=1)', type='l')
```



- Notes about LP:
 - for a given problem the range may be much narrower than $[-4, 4]$
 - when any predictor X_j with a non-zero β_j is categorical, LP cannot take on all possible values within its range, so the above plot will have gaps
- As a special case the model can estimate and compare two probabilities as done above, through an odds ratio
- Logistic regression seems like an overkill here, but it sets the stage for more complex frequentist analysis as well as Bayesian analysis
- For our case the model is as follows
- Consider groups A and B, A = reference group
 $[x]$ denotes 1 if x is true, 0 if x is false
 Define the expit function as the inverse of the logit function, or $\text{expit}(x) = \frac{1}{1+\exp(-x)}$

$$P(Y = 1|\text{group}) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1[\text{group } B]))}$$

$$\begin{aligned}
 &= \text{expit}(\beta_0 + \beta_1[\text{group } B]) \\
 P(Y = 1|\text{group A}) &= \text{expit}(\beta_0) \\
 P(Y = 1|\text{group B}) &= \text{expit}(\beta_0 + \beta_1)
 \end{aligned}$$

$\beta_0 = \log \text{ odds of probability of event in group A} = \log\left(\frac{p_1}{1-p_1}\right) = \text{logit}(p_1)$

$\beta_1 = \text{increase in log odds in going from group A to group B} =$

$\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right) = \text{logit}(p_2) - \text{logit}(p_1)$

$\text{exp}(\beta_1) = \text{group B : group A odds ratio} = \frac{\frac{p_2}{1-p_2}}{\frac{p_1}{1-p_1}}$

$\text{expit}(\beta_0) = p_1$

$\text{expit}(\beta_0 + \beta_1) = p_2$

- Once the β s are estimated, one quickly gets the B:A odds ratio and \hat{p}_1 and \hat{p}_2
- This model is **saturated**
 - has the maximum number of parameters needed to fully describe the situation (here, 2 since 2 groups)
 - saturated models **must** fit the data if the distributional and independence assumptions are met
 - logistic model has no distributional assumption
- Logistic regression is more general and flexible than the specialized tests for proportions
 - allows testing association on continuous characteristics
 - easily extends to more than two groups
 - allows adjustment for covariates
- Examples:
 - assess effects of subjects' sex and country (Canada vs. US) on $P(Y = 1)$ denoted by p
 - $\text{logit } p = \text{constant} + \text{logit male effect} + \text{logit Canada effect}$
 - same but allow for interaction

– assess effects of subjects' sex and country (Canada vs. US) on $P(Y = 1)$ denoted by p

$\text{logit } p = \text{constant} + \text{logit male effect} + \text{logit Canada effect}$

– same but allow for interaction

$\text{logit } p = \text{constant} + \text{logit male effect} + \text{logit Canada effect} + \text{special effect of being male if Canadian}$

- latter model is saturated with 3 d.f. so fits as well as a model with 4 proportions
 - * unlike the overall Pearson χ^2 test, allows testing interaction and
 - * separate effects of sex and country (e.g., 2 d.f. chunk test for whether there is a sex difference for either country, allowing for the sex effect to differ by country)

6.10.1

Test Statistics

For frequentist logistic models there are 3 types of χ^2 test statistics for testing the same hypothesis:

- likelihood ratio (LR) test (usually the most accurate) and is scale invariant^d
 - Can obtain the likelihood ratio χ^2 statistic from either the logistic model or from a logarithmic equation in the two proportions and sample sizes
- Score test (identical to Pearson test for overall model if model is saturated)
- Wald test (square of $\frac{\hat{\beta}}{\text{s.e.}}$ in the one parameter case; misbehaves for extremely large effects)
- Wald test is the easiest to compute but P -values and confidence intervals from it are not as accurate. The score test is the way to exactly reproduce the Pearson χ^2 statistic from the logistic model.
- As with t -test vs. a linear model, the special case tests are not needed once you use the logistic model framework
- Three usages of any of these test statistics:
 - individual test, e.g. sex effect in sex-country model without interaction

^dThe LR test statistic is the same whether testing for an absolute risk difference of 0.0, or for an odds or risk ratio of 1.0.

- chunk test, e.g. sex + sex × country interaction 2 d.f. test tests overall sex effect
- global test of no association, e.g. 3 d.f. test for whether sex or country is associated with Y

6.10.2

Frequentist Analysis Example

- Consider again our emergency surgery example
- String the observations out to get one row = one patient, binary Y

```
require(rms)
options(prType='latex')
priority <- factor(c(rep('emergency', 25), rep('other', 111)), c('other', 'emergency'))
death <- c(rep(0, 19), rep(1, 6), rep(0, 100), rep(1, 11))
table(priority, death)
```

	death	
priority	0	1
other	100	11
emergency	19	6

```
d <- data.frame(priority, death)
dd <- datadist(d); options(datadist='dd')
# rms package needs covariate summaries computed by datadist
f <- lrm(death ~ priority)
f
```

Logistic Regression Model

```
lrm(formula = death ~ priority)
```

		Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs	136	LR χ^2	R^2	C 0.597
0	119	d.f.	g	D_{xy} 0.193
1	17	$\Pr(> \chi^2)$	g_r	γ 0.483
$\max \frac{\partial \log L}{\partial \beta} 4 \times 10^{-9}$			g_p	τ_a 0.043
			Brier	0.106

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
Intercept	-2.2073	0.3177	-6.95	<0.0001
priority=emergency	1.0546	0.5659	1.86	0.0624

- Compare the LR χ^2 of 3.2 with the earlier Pearson χ^2 of 3.70
- The likelihood ratio (LR) χ^2 test statistic and its P -value are usually a little more accurate than the other association tests
 - but χ^2 distribution is still only an approximation to the true sampling distribution
- See that we can recover the simple proportions from the fitted logistic model:

$$\hat{p}_1 = \frac{11}{111} = 0.099 = \text{expit}(-2.2073)$$

$$\hat{p}_2 = \frac{6}{25} = 0.24 = \text{expit}(-2.2073 + 1.0546)$$

```
summary(f)
```

	Low	High	Δ	Effect	S.E.	Lower	0.95	Upper	0.95
priority — emergency:other	1	2		1.0546	0.56587	-0.054487		2.1637	
Odds Ratio	1	2		2.8708		0.946970		8.7031	

- The point estimate and CLs for the odds ratio is the same as what we obtained earlier.

Add a random binary variable to the logistic model—one that is correlated with the surgical priority—to see the effect on the estimate of the priority effect

```
set.seed(10)
randomUniform <- runif(length(priority))
random <- ifelse(priority == 'emergency', randomUniform < 1/3,
                  randomUniform > 1/3) * 1
d$random <- random
table(priority, random)
```

	random
priority	0 1
other	39 72
emergency	17 8

```
lrm(death ~ priority + random)
```

Logistic Regression Model

```
lrm(formula = death ~ priority + random)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	136	LR χ^2	6.03	R^2	0.082	C	0.659
0	119	d.f.	2	g	0.638	D_{xy}	0.319
1	17	$\Pr(>\chi^2)$	0.0489	g_r	1.892	γ	0.464
$\max \left \frac{\partial \log L}{\partial \beta} \right 1 \times 10^{-8}$				g_p	0.073	τ_a	0.070
				Brier	0.103		

	$\hat{\beta}$	S.E.	Wald Z	$\Pr(> Z)$
Intercept	-1.6851	0.4157	-4.05	<0.0001
priority=emergency	0.7811	0.5909	1.32	0.1862
random	-0.9319	0.5621	-1.66	0.0973

The effect of emergency status is diminished, and the random grouping variable, created to have no relation to death in the population, has a large apparent effect.

6.10.3

Bayesian Logistic Regression Analysis

- Has several advantages
 - All calculations are exact (to within simulation error) without changing the model
 - Can incorporate external information
 - Intuitive measures of evidence
 - Automatically handles zero-frequency cells (priors shrink probabilities a bit away from 0.0)
- Simple to use beta priors for each of the two probabilities

- But we'd need to incorporate a complex dependency in the two priors because we know more about how the two probabilities relate to each other than we know about each absolute risk
- Simpler to have a wide prior on p_1 and to have a non-flat prior on the log odds ratio
 - could backsolve and show dependency between knowledge of p_1 and p_2
- Use the same data model as above
- As before we use the R `brms` package which makes standard modeling easy
- Need two priors: for intercept β_0 and for log odds ratio β_1

- β_0 : use a normal distribution that makes $p_1 = 0.05$ the most likely value (put mean at $\text{logit}(0.05) = -2.944$) and allows only a 0.1 chance that $p_1 > 0.2$; solve for SD σ that accomplishes that

```
# Given mu and value, solve for SD so that the tail area of the normal
# distribution beyond value is prob
normsolve ← function(mu, value, prob) (value - mu) / qnorm(1 - prob)
normsolve(qlogis(0.05), qlogis(0.2), 0.1) # qlogis is R logit()
```

```
[1] 1.215827
```

- We round σ to 1.216
- For β_1 put a prior that has equal chance for $\text{OR} < 1$ as for $\text{OR} > 1$, i.e., mean for log OR of zero
Put a chance of only 0.1 that $\text{OR} > 3$

```
normsolve(0, log(3), 0.1)
```

```
[1] 0.8572517
```

- Round to 0.857

Compute the correlation between prior evidence for p_1 and p_2 by drawing 100,000 samples from the prior distributions. Also verify that prior probability $p_2 > p_1$ is $\frac{1}{2} = \text{prob. } \text{OR} > 1$.

```
b0 ← rnorm(100000, qlogis(0.05), 1.216)
b1 ← rnorm(100000, 0, 0.857)
p1 ← plogis(b0)
p2 ← plogis(b0 + b1)
cor(b0, b1, method='spearman')
```

```
[1] 0.003724938
```

```
cor(p1, p2, method='spearman')
```

```
[1] 0.8066422
```

```
# Define functions for posterior probability operator and posterior mode
P ← mean # proportion of posterior draws for which a condition holds
pmode ← function(x) {
  z ← density(x)
  z$x[which.max(z$y)]
}
```

```
P(p2 > p1)
```

```
[1] 0.50106
```

```
P(b1 > 0)
```

```
[1] 0.50106
```

```
P(exp(b1) > 1)
```

```
[1] 0.50106
```

To show that prior knowledge about p_1 and p_2 is uncorrelated when we don't know anything about the odds ratio, repeat the above calculation use a SD of 1000 for the log odds ratio:

```
b1 ← rnorm(100000, 0, 1000)
p2 ← plogis(b0 + b1)
cor(p1, p2, method='spearman')
```

```
[1] 0.001448724
```

Now do Bayesian logistic regression analysis.

```
require(brms)
# Tell brms/Stan to use all available CPU cores
options(mc.cores=parallel::detectCores())
```

```
p ← c(prior(normal(-2.944, 1.216), class='Intercept'),
      prior(normal(0, 0.857), class='b'))
f ← brm(death ~ priority, data=d, prior=p, family='bernoulli', seed=123)
```

f

```

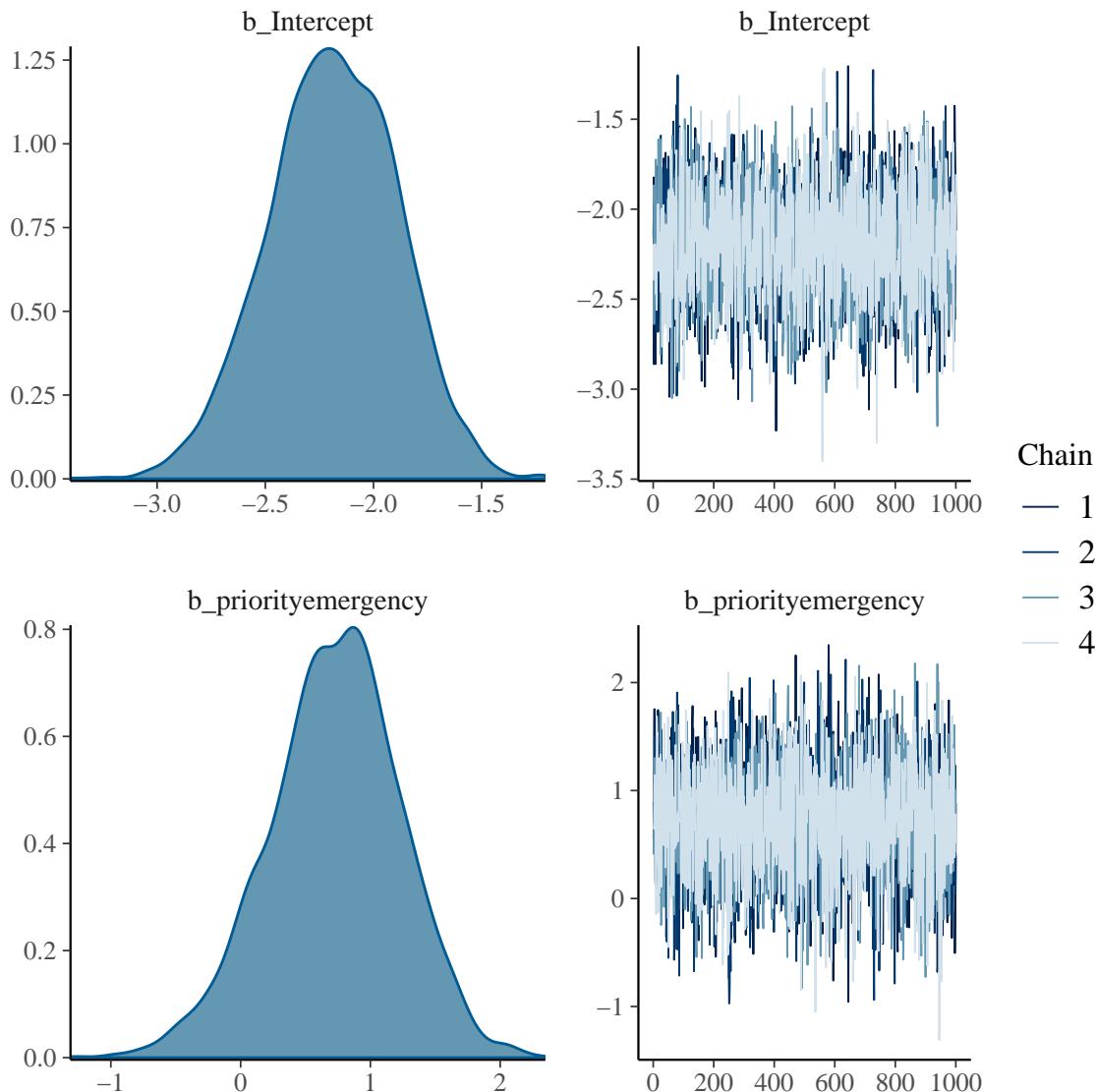
Family: bernoulli
Links: mu = logit
Formula: death ~ priority
Data: d (Number of observations: 136)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
Estimate  Est.Error  l-95% CI  u-95% CI   Rhat Bulk_ESS Tail_ESS
Intercept     -2.19      0.30    -2.78    -1.63 1.00     2789     2493
priorityemergency  0.72      0.51    -0.33     1.66 1.00     2655     2181

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

```

plot(f)



```
# Bring out posterior draws
w <- as.data.frame(f)
b0 <- w[, 'b_Intercept']
b1 <- w[, 'b_priorityemergency']
r <- rbind(c(mean(b0), median(b0), pmode(b0)),
            c(mean(b1), median(b1), pmode(b1)),
            c(mean(exp(b1)), median(exp(b1)), pmode(exp(b1))))
colnames(r) <- c('Posterior Mean', 'Posterior Median', 'Posterior Mode')
rownames(r) <- c('b0', 'b1', 'OR')
round(r, 3)
```

	Posterior Mean	Posterior Median	Posterior Mode
b0	-2.188	-2.184	-2.203
b1	0.724	0.748	0.865
OR	2.333	2.113	1.721

Because the prior on the OR is conservative, the Bayesian posterior mode for the OR is smaller than the frequentist maximum likelihood estimate of 2.87.

Below notice how easy it is to do Bayesian inference on derived quantities p1 and p2 which are functions of b0 and b1.

```
# 0.95 credible interval for log odds ratio and odds ratio
quantile(b1, c(0.025, 0.975))
```

```
2.5%      97.5%
-0.326236  1.656346
```

```
quantile(exp(b1), c(.025, 0.975))
```

```
2.5%      97.5%
0.721635  5.240131
```

```
exp(quantile(b1, c(0.025, 0.975)))
```

```
2.5%      97.5%
0.7216349 5.2401303
```

```
# Posterior density of emergency:other odds ratio
plot(density(exp(b1)), xlab='OR', main='')
abline(v=c(1, pmode(exp(b1))), col=gray(0.85))
# Probability that OR > 1
P(exp(b1) > 1)
```

```
[1] 0.91875
```

```
# Probability it is > 1.5
P(exp(b1) > 1.5)
```

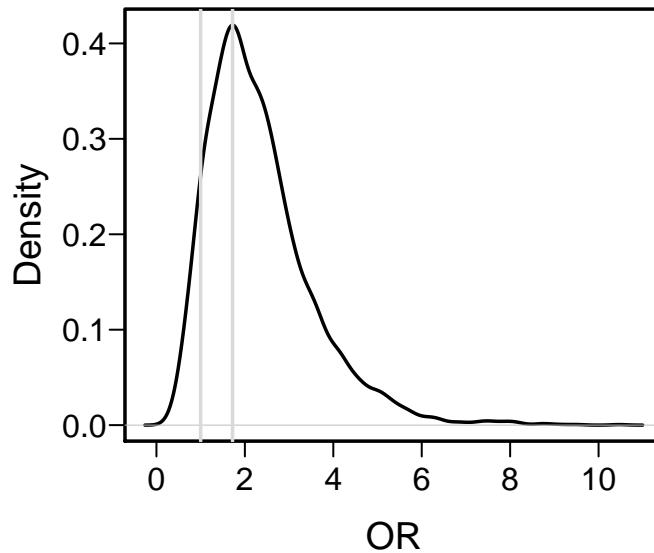
```
[1] 0.74925
```

```
# Probability that risk with emergency surgery exceeds that of
# non-emergency (same as P(OR > 1))
# plogis in R is 1/(1 + exp(-x))
P(plogis(b0 + b1) > plogis(b0))
```

```
[1] 0.91875
```

```
# Prob. that risk with emergency surgery elevated by more than 0.03
P(plogis(b0 + b1) > plogis(b0) + 0.03)
```

```
[1] 0.8165
```



Even though the priors for the intercept and log odds ratio are independent, the connection of these two parameters in the data likelihood makes the posteriors dependent as shown with Spearman correlations of the posterior draws below. Also get the correlation between evidence for the two probabilities. These have correlated priors even though they are unconnected in the likelihood function. Posteriors for p_1 and p_2 are less correlated than their priors.

```
cor(b0, b1, method='spearman')
```

```
[1] -0.4439781
```

```
cor(plogis(b0), plogis(b0 + b1), method='spearman')
```

```
[1] 0.1385611
```

To demonstrate the effect of a skeptical prior:

- Add random grouping to model as we did with the frequentist analysis

- Make use of prior information that this variable is unlikely to be important
- Put a prior on the log OR for this variable centered at zero with chance that the OR > 1.25 of only 0.05

```
normsolve(0, log(1.25), 0.05)
```

```
[1] 0.1356616
```

```
p ← c(prior(normal(-2.944, 1.216), class='Intercept'),
       prior(normal(0, 0.857), class='b', coef='priorityemergency'),
       prior(normal(0, 0.136), class='b', coef='random'))

f2 ← brm(death ~ priority + random, data=d, prior=p, family='bernoulli',
          seed=121, refresh=FALSE)
```

```
f2
```

```
Family: bernoulli
Links: mu = logit
Formula: death ~ priority + random
Data: d (Number of observations: 136)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
              Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      -2.16     0.31    -2.82   -1.61 1.00    2804    2922
priorityemergency 0.71     0.50    -0.27    1.66 1.00    3614    3020
random        -0.06     0.13    -0.31    0.19 1.00    3879    3110
```

Samples were drawn using sampling(NUTS). For each parameter, Eff.Sample is a crude measure of effective sample size, and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat = 1).

- Effect of random is greatly discounted
- Posterior mean priority effect and its credible interval is virtually the same as the model that excluded random

James Rae and Nils Reimer have written a nice tutorial on using the R `brms` package for binary logistic regression available at bit.ly/brms-lrm

Chapter 7

Nonparametric Statistical Tests



7.1

When to use non-parametric methods

- Short answer: Good default when P -values are needed and there are no covariates to adjust for
- Nonparametric methods are those not requiring one to assume a certain distribution for the raw data
 - In contrast, parametric methods assume data come from some underlying distribution
 - t -tests assume the data come form a Gaussian distribution
- Response variable ordinal or interval
- For ordinal responses nonparametric methods are preferred because they assume no spacing between categories
- No problem in using nonparametric tests on interval data
 - if normality holds, nonpar. test 0.95 efficient, i.e., has about the same power as the parametric test done on 0.95 of the observations^a
 - if normality does not hold, nonpar. tests can be arbitrarily more efficient and

^aThe large-sample efficiency of the Wilcoxon and Spearman tests compared to t and r tests is $\frac{3}{\pi} = 0.9549$.

- powerful than the corresponding parametric test
- an elegant and non-arbitrary way to deal with extreme values or outliers
 - rank-based nonparametric tests give the analyst freedom from having to choose the correct transformation of the measurement (as long as the optimum transformation is monotonic)
- Nonparametric methods are robust, many parametric methods are not
 - Example: *t*-test comparing two sets of measurements
1 2 3 4 5 6 7 8 9 10 vs. 7 8 9 10 11 12 13 14 15 16 17 18 19 20
means: 5.5 and 13.5, $P = 0.000019$
1 2 3 4 5 6 7 8 9 10 vs. 7 8 9 10 11 12 13 14 15 16 17 18 19 20 **200**
means: 5.5 and 25.9, $P = 0.12$
The SD is a particularly non-robust statistical estimator.
 - Example: Fecal calprotectin being evaluated as a possible biomarker of disease severity (Figure 7.1)
 - Calprotectin has an upper detection limit
 - Median can be calculated (mean cannot)
 - If all you want is a P -value nonpar. tests are preferred
 - Especially if response is univariate and no need to adjust for covariates
 - Pre-testing for normality and deciding nonparametric vs. parametric analysis is a bad idea
 - Tests for normality do not have a power of 1.0 and type I error of 0.0
 - Leads to temptation, e.g., an investigator might “forget” to do the test of normality if the *t*-test is significant
 - Doesn’t acknowledge that nonparametric tests are very efficient even under normality
 - Pre-testing for normality alters the type I error and confidence interval coverage

- A drawback is that nonpar. tests do not correspond to usual confidence limits for effects
 - E.g., a CL for the difference in 2 means may include zero whereas the Wilcoxon test yields $P = 0.01$
 - Point estimate that exactly corresponds to the Wilcoxon two-sample test is the Hodges-Lehman estimate of the location difference
 - * median of all possible differences between a measurement from group 1 and a measurement from group 2
- Nonparametric tests are often obtained by replacing the data with ranks across subjects and then doing the parametric test
- Many nonpar. tests give the same P -value regardless of how the data are transformed; a careful choice of transformation (e.g., log) must sometimes be used in the context of parametric tests
- P -values computed using e.g. the t distribution are quite accurate for nonparametric tests
- In case of ties, midranks are used, e.g., if the raw data were 105 120 120 121 the ranks would be 1 2.5 2.5 4

Parametric Test	Nonparametric Counterpart	Semiparametric Model Counterpart
1-sample t	Wilcoxon signed-rank	
2-sample t	Wilcoxon 2-sample rank-sum	Proportional odds
k -sample ANOVA	Kruskal-Wallis	Proportional odds
Pearson r	Spearman ρ	

7.2

One Sample Test: Wilcoxon Signed-Rank

- Almost always used on paired data where the column of values represents differences (e.g., post-pre) or log ratios
- The *sign test* is the simplest test for the median difference being zero in the population
 - it just counts the number of positive differences after tossing out zero differences
 - tests $H_0 : \text{Prob}[x > 0] = \frac{1}{2}$, i.e., that it is equally likely in the population to have a value below zero as it is to have a value above zero
 - as it ignores magnitudes completely, the test is inefficient
- By contrast, with the much more powerful Wilcoxon signed rank one-sample test, ranks of absolute differences are given the sign of the original difference
- Magnitudes of raw data matter more here than with the Wilcoxon 2-sample test
- Example: A crossover study in which the treatment order is randomized
Data arranged so that treatment A is in the first column, no matter which order treatment A was given

A	B	B-A	Rank	$ B - A $	Signed Rank
5	6	1	1.5	1.5	1.5
6	5	-1	1.5	1.5	-1.5
4	9	5	4.0	4.0	4.0
7	9	2	3.0	3.0	3.0

- A good approximation to an exact P -value may be obtained by computing

$$z = \frac{\sum SR_i}{\sqrt{\sum SR_i^2}},$$

where the signed rank for observation i is SR_i . This formula already takes ties into account without using Rosner's messy Eq. 9.5. We look up $|z|$ against the normal distribution. Here $z = \frac{7}{\sqrt{29.5}} = 1.29$ and the 2-tailed P -value is given below.

```

sr ← c(1.5, -1.5, 4, 3)
z ← sum(sr) / sqrt(sum(sr ^ 2))
pval ← 2 * (1 - pnorm(abs(z)))
c(z=z, pval=pval)

```

	<i>z</i>	<i>pval</i>
	1.2888045	0.1974661

- If all differences are positive or all are negative, the exact 2-tailed *P*-value is $\frac{1}{2^{n-1}}$
 - implies that *n* must exceed 5 for any possibility of significance at the $\alpha = 0.05$ level for a 2-tailed test

7.2.1

One sample/Paired Test Example

- Sleep Dataset
 - Compare the effects of two soporific drugs.
 - Each subject receives placebo, Drug 1, and Drug 2
 - Study question: Is Drug 1 or Drug 2 more effective at increasing sleep?
 - Dependent variable: Difference in hours of sleep comparing Drug 2 to Drug 1
 - H_0 : For any given subject, the difference in hours of sleep is equally likely to be positive or negative
 - See P. 5-79 for a parametric test on these data

```

drug1 ← c(.7, -1.6, -.2, -1.2, -.1, 3.4, 3.7, .8, 0, 2)
drug2 ← c(1.9, .8, 1.1, .1, -.1, 4.4, 5.5, 1.6, 4.6, 3.4)
wilcox.test(drug2, drug1, paired=TRUE)

```

```

Wilcoxon signed rank test with continuity correction

data: drug2 and drug1
V = 45, p-value = 0.009091
alternative hypothesis: true location shift is not equal to 0

```

```
wilcox.test(drug2 - drug1)
```

```
Wilcoxon signed rank test with continuity correction
```

```
data: drug2 - drug1
V = 45, p-value = 0.009091
alternative hypothesis: true location is not equal to 0
```

```
wilcox.test(drug2 - drug1, correct=FALSE)
```

```
Wilcoxon signed rank test
```

```
data: drug2 - drug1
V = 45, p-value = 0.007632
alternative hypothesis: true location is not equal to 0
```

```
sr ← c(3, 8, 4.5, 4.5, 0, 2, 7, 1, 9, 6)
z ← sum(sr) / sqrt(sum(sr ^ 2))
c(z=z, pval=2 * (1 - pnorm(abs(z))))
```

```
z          pval
2.667911250 0.007632442
```

```
d ← data.frame(Drug=c(rep('Drug 1', 10), rep('Drug 2', 10),
                      rep('Difference', 10)),
                  extra=c(drug1, drug2, drug2 - drug1))
```

- Interpretation: Reject H_0 , Drug 2 increases sleep by the same hours as Drug 1 ($p = 0.008$)
- Could also perform sign test on sleep data
 - If drugs are equally effective, should have same number of '+' and '-'
 - Observed data: 0 '−', 9 '+', throw out 1 'no change'

Subject	Drug 1	Drug 2	Diff (2-1)	Sign	Rank
1	0.7	1.9	1.2	+	3
2	-1.6	0.8	2.4	+	8
3	-0.2	1.1	1.3	+	4.5
4	-1.2	0.1	1.3	+	4.5
5	-0.1	-0.1	0.0	NA	
6	3.4	4.4	1.0	+	2
7	3.7	5.5	1.8	+	7
8	0.8	1.6	0.8	+	1
9	0.0	4.6	4.6	+	9
10	2.0	3.4	1.4	+	6

Table 7.1: Hours of extra sleep on drugs 1 and 2, differences, signs and signed ranks of sleep study data

- Sign test (2-sided) P -value: Probability of observing 9 of 9 + or 9 of 9 -
- $p = 0.004$, so evidence against H_0

```
2 * (1 / 2) ^ 9      # 2 * to make it two-tailed
[1] 0.00390625
```

- The signed rank test assumes that the distribution of differences is symmetric
- It tests whether the median difference is zero
- Also tests that the mean is zero
- In general it tests that, for two randomly chosen observations i and j with values (differences) x_i and x_j , that the probability that $x_i + x_j > 0$ is $\frac{1}{2}$
- The estimator that corresponds exactly to the test in all situations is the pseudomedian, the median of all possible pairwise averages of x_i and x_j , so one could say that the signed rank test tests H_0 : pseudomedian=0
- The value $\frac{\overline{SR}}{n+1} - \frac{1}{2}$ estimates the probability that two randomly chosen observations have a positive sum, where \overline{SR} is the mean of the column of signed ranks
- To test $H_0 : \eta = \eta_0$, where η is the population median (not a difference) and η_0 is some constant, we create the n values $x_i - \eta_0$ and feed those to the signed rank test, assuming the distribution is symmetric
- When all nonzero values are of the same sign, the test reduces to the *sign test* and the 2-tailed P -value is $(\frac{1}{2})^{n-1}$ where n is the number of nonzero values

Test whether the continuity correction makes P -values closer to the exact calculation^b, and compare to our simple formula.

```
# Assume we are already starting with signed ranks as x
wsr ← function(x, ...) wilcox.test(x, ...)$p.value
sim ← function(x) {
  z ← sum(x) / sqrt(sum(x ^ 2))
  2 * (1 - pnorm(abs(z)))
}
all ← function(x) round(c(
  continuity=wsr(x, correct=TRUE, exact=FALSE),
```

^bThe exact P -value is available only when there are no ties.

```
nocontinuity=wsr(x, correct=FALSE, exact=FALSE),
exact=wsr(x, exact=TRUE),
simple=sim(x)), 4)
all(1:4)
```

continuity	nocontinuity	exact	simple
0.1003	0.0679	0.1250	0.0679

```
all(c(-1, 2 : 4))
```

continuity	nocontinuity	exact	simple
0.2012	0.1441	0.2500	0.1441

```
all(c(-2, c(1, 3, 4)))
```

continuity	nocontinuity	exact	simple
0.3613	0.2733	0.3750	0.2733

```
all(c(-1, -2, 3 : 5))
```

continuity	nocontinuity	exact	simple
0.2807	0.2249	0.3125	0.2249

```
all(c(-5, -1, 2, 3, 4, 6))
```

continuity	nocontinuity	exact	simple
0.4017	0.3454	0.4375	0.3454

From these examples the guidance is to:

- Use the exact calculation if there are no ties
- Otherwise use the continuity correction (i.e., the default in `wilcox.test`) unlike the recommendation for the Pearson χ^2 test

7.3

Two Sample Test: Wilcoxon–Mann–Whitney

- The Wilcoxon–Mann–Whitney (WMW) 2-sample rank sum test is for testing for equality of central tendency of two distributions (for unpaired data)
- Ranking is done by combining the two samples and ignoring which sample each observation came from
- Example:

Females	120	118	121	119
Males	124	120	133	
Ranks for Females	3.5	1	5	2
Ranks for Males	6	3.5	7	

- Doing a 2-sample t -test using these ranks as if they were raw data and computing the P -value against $4+3-2=5$ d.f. will work quite well
- Some statistical packages compute P -values exactly (especially if there are no ties)
- Loosely speaking the WMW test tests whether the population medians of the two groups are the same
- More accurately and more generally, it tests whether observations in one population tend to be larger than observations in the other
- Letting x_1 and x_2 respectively be randomly chosen observations from populations one and two, WMW tests $H_0 : c = \frac{1}{2}$, where $c = \text{Prob}[x_1 > x_2]$
- The c index (*concordance probability*) may be estimated by computing

$$c = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2},$$

where \bar{R} is the mean of the ranks in group 1;

For the above data $\bar{R} = 2.875$ and $c = \frac{2.875-2.5}{3} = 0.125$, so we estimate that the probability is 0.125 that a randomly chosen female has a value greater than a randomly chosen male.

- In diagnostic studies where x is the continuous result of a medical test and the grouping variable is diseased vs. non-diseased, c is the area under the receiver operating characteristic (ROC) curve
- Test still has the “probability of ordering” interpretation when the variances of the two samples are markedly different, but it no longer tests anything like the difference in population medians

If there is no overlap between measurements from group 1 and those from group 2, the exact 2-sided P -value for the Wilcoxon test is $2/\frac{n!}{n_1!n_2!}$. If $n_1 = n_2$, n_1 must be ≥ 4 to obtain $P < 0.05$ (in this case $P = 0.029$).

7.3.1

Two Sample WMW example

- Fecal calprotectin being evaluated as a possible biomarker of disease severity
- Calprotectin measured in 26 subjects, 8 observed to have no/mild activity by endoscopy
- Calprotectin has upper detection limit at 2500 units
 - A type of missing data, but need to keep in analysis
- Study question: Are calprotectin levels different in subjects with no or mild activity compared to subjects with moderate or severe activity?
- Statement of the null hypothesis
 - H_0 : Populations with no/mild activity have the same distribution of calprotectin as populations with moderate/severe activity
 - $H_0 : c = \frac{1}{2}$

```
# Fecal Calprotectin: 2500 is above detection limit
calpro ← c(2500, 244, 2500, 726, 86, 2500, 61, 392, 2500, 114, 1226,
       2500, 168, 910, 627, 2500, 781, 57, 483, 30, 925, 1027,
       2500, 2500, 38, 18)

# Endoscopy score: 1 = No/Mild, 2=Mod/Severe Disease
# Would have been far better to code dose as 4 ordinal levels
```

```

endo <- c(2, 1, 2, 2, 1, 1, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2, 1, 2,
        2, 2, 2, 2, 1, 1)
endo <- factor(endo, 1 : 2,
               c("No or Mild Activity", "Moderate or Severe Activity"))
require(ggplot2) # Fig. 7.1
ggplot(data.frame(endo, calpro), aes(y=calpro, x=endo)) +
  geom_boxplot(color='lightblue', alpha=.85, width=.4) +
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
  xlab('') + ylab('Fecal Calprotectin') + coord_flip() +
  geom_hline(aes(yintercept=2500, col=I('red')), linetype='dotted')

wilcox.test(calpro ~ endo)

```

```

Wilcoxon rank sum test with continuity correction

data: calpro by endo
W = 23.5, p-value = 0.006814
alternative hypothesis: true location shift is not equal to 0

```

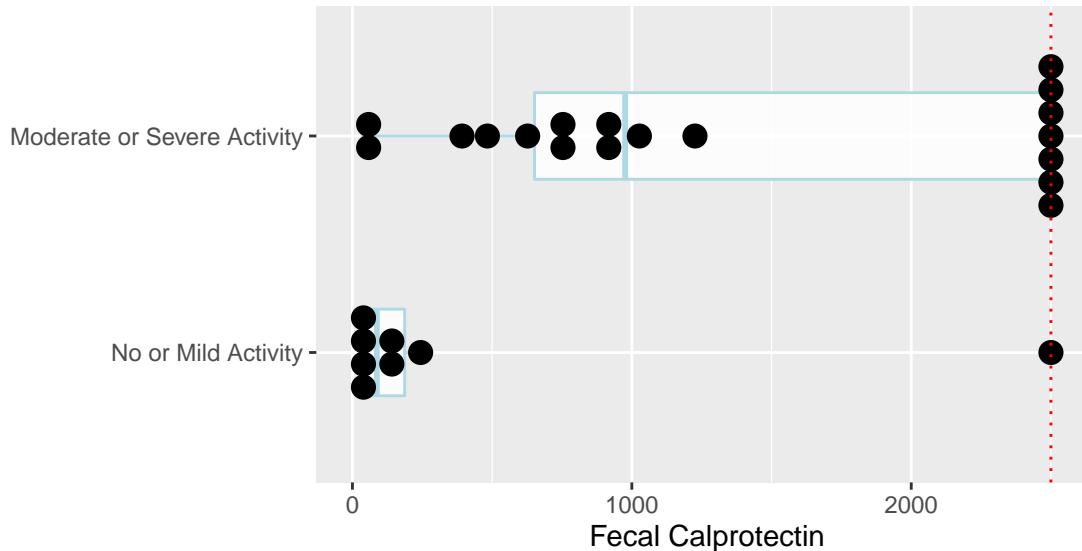


Figure 7.1: Fecal calprotectin by endoscopy severity rating. Red dotted line is the detection limit. Ordinal disease categories should not have been combined.

The following plots the ranks that are used in the Wilcoxon-Mann-Whitney two-sample rank sum test.

```

ggplot(data.frame(endo, calpro), aes(y=rank(calpro), x=endo)) + # Fig 7.2
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
  xlab('') + ylab('Rank of Fecal Calprotectin') + coord_flip()

```

- Test statistic W equals the sum of the ranks in the no/mild group minus $n_1 * (n_1 + 1)/2$, where n_1 is the number of subjects in the no/mild sample
- $W = 59.5 - \frac{8*9}{2} = 23.5$

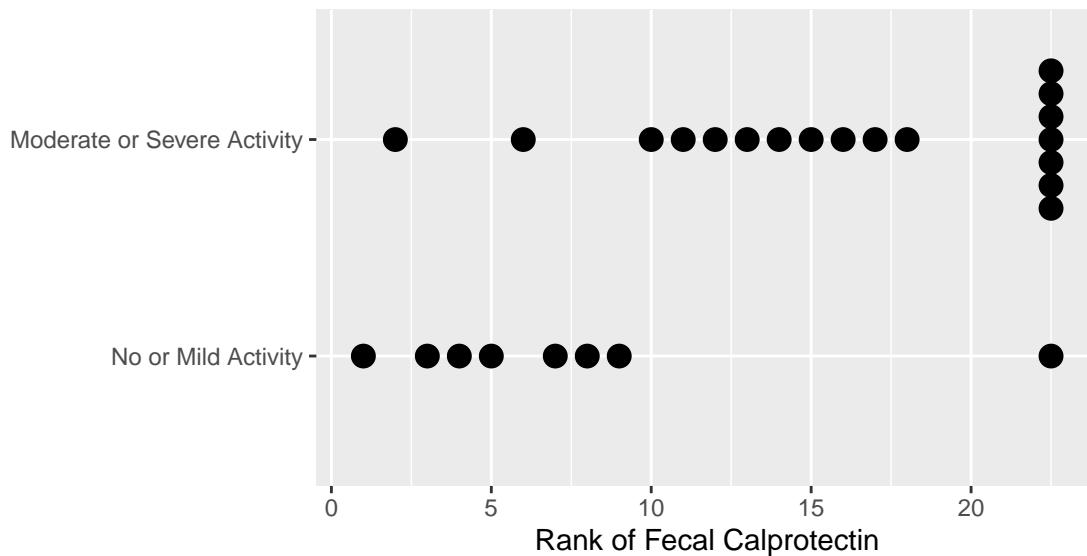


Figure 7.2: Ranks of calprotectin

- A common (but loose) interpretation: People with moderate/severe activity have higher *median* fecal calprotectin levels than people with no/mild activity ($p = 0.007$).
- Better: remove *median* and supplement with the *c*-index (concordance probability) or Somers' D_{xy} rank correlation between calprotectin and endoscopy status. The code for the R `somers2` function shows how the concordance probability is computed from the mean of the ranks in one of the two groups.

```
require(Hmisc)
# Convert endo to a binary variable
somers2(calpro, endo=='Moderate or Severe Activity')
```

C	Dxy	n	Missing
0.8368056	0.6736111	26.0000000	0.0000000

If you type `somers2` to list the code for the function you will see that the *c*-index is tightly related to the Wilcoxon test when you see this code:

```
mean.rank <- mean(rank(x)[y == 1])
c.index <- (mean.rank - (n1 + 1)/2) / (n - n1)
```

7.3.2

Point and Interval Estimates for Wilcoxon Two-Sample Comparison

As mentioned earlier, the effect estimate that is exactly consistent with the Wilcoxon two-sample test is the robust Hodges-Lehman estimator—the median of all possible differences between a measurement from group 1 and a measurement from group 2. There is a confidence interval for this estimator.

- Assume data come from distributions with same shape and differ only in location
- Consider a sample of 4 males and 3 females
- Difference in sample medians is $124 - 119.5 = 4.5$
- Consider all possible differences between sample 1 and sample 2

Male	Female			
	120	118	121	119
124	4	6	3	5
120	0	2	-1	1
133	13	15	12	14

- Hodges-Lehman estimate of the sex effect: median of the 12 differences = 4.5
- In this case equaled difference in sample medians just by coincidence

```
female <- c(120, 118, 121, 119)
male   <- c(124, 120, 133)
differences <- outer(male, female, '-')
differences
```

```
[ ,1] [ ,2] [ ,3] [ ,4]
[1,]    4     6     3     5
[2,]    0     2    -1     1
[3,]   13    15    12    14
```

```
median(differences)
```

```
[1] 4.5
```

```
# Can't figure out how difference in location is computed below
# It's not the Hodges-Lehman estimate
wilcox.test(male, female, conf.int=TRUE)
```

```
Wilcoxon rank sum test with continuity correction

data: male and female
W = 10.5, p-value = 0.1536
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1 15
sample estimates:
difference in location
        4.791134
```

In general, $1 - \alpha$ confidence intervals are the set of values that if hypothesized to be the true location parameter would not be rejected at the α level. `wilcox.test` computes the location shift by solving for the hypothesized value that yields $P = 1.0$ instead of the more proper median of all differences. Look into this further by plotting the P -value as a function of the hypothesized value.

```
dif  ← seq(-3, 15, by=.1)
n    ← length(dif)
pval ← numeric(n)
for(i in 1 : n) pval[i] ← wilcox.test(male - dif[i], female)$p.value

ggplot(data.frame(dif, pval), aes(x=dif, y=pval)) +
  geom_step() +
  geom_hline(yintercept=.05, col='red', linetype='dotted') +
  geom_vline(xintercept=c(4.5, 4.791, -1, 15), col='red', linetype='dotted') +
  xlab('Difference') + ylab('P-value')
```

See Section 7.4 for a more approximate confidence interval.

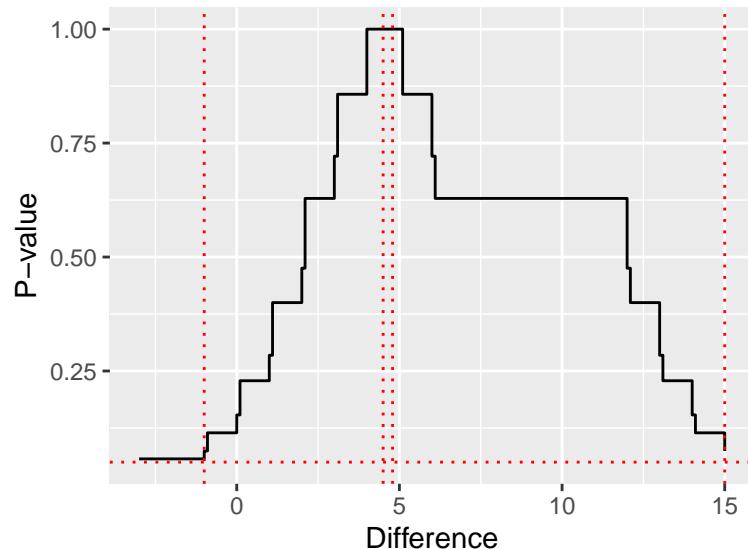


Figure 7.3: Wilcoxon P -value vs. hypothesized male-female difference. Horizontal line is $P = 0.05$. Vertical lines from left to right are the lower 0.95 confidence limit from `wilcox.test`, the median difference, the Hodges-Lehman estimator as computed by `wilcox.test`, and the upper 0.95 confidence limit from `wilcox.test`.

7.4

Confidence Intervals for Medians and Their Differences

- Confidence intervals for the median (one sample)
 - Table 18.4 (Altman) gives the ranks of the observations to be used to give approximate confidence intervals for the median
 - e.g., if $n = 12$, the 3rd and 10th largest values give a 0.961 confidence interval
 - For larger sample sizes, the lower ranked value (r) and upper ranked value (s) to select for an approximate 0.95 confidence interval for the population median is
- e.g., if $n = 100$ then $r = 40.2$ and $s = 60.8$, so we would pick the 40th and 61st largest values from the sample to specify a 0.95 confidence interval for the population median
- For exact confidence interval for the median see stats.stackexchange.com/questions/186957, which also discusses why there is no exact nonparametric confidence interval for the mean. Let's get the exact order statistics that result in an exact confidence interval for the median:

```
# Exact CI for median from DescTools package SignTest.default
# See also ttp://www.stat.umn.edu/geyer/old03/5102/notes/rank.pdf,
# http://describd.com/doc/75941305/
# Confidence-Interval-for-Median-Based-on-Sign-Test
cimed <- function(x, alpha=0.05, na.rm=FALSE) {
  if(na.rm) x <- x[! is.na(x)]
  n <- length(x)
  k <- qbinom(p=alpha / 2, size=n, prob=0.5, lower.tail=TRUE)
  ## Actual CL: 1 - 2 * pbinom(k - 1, size=n, prob=0.5) ≥ 1 - alpha
  sort(x)[c(k, n - k + 1)]
}
```

```
[1] 40 61
```

For $n = 100$ we see that the approximate interval happened to be exact.

- Confidence intervals for the difference in two medians (two samples)

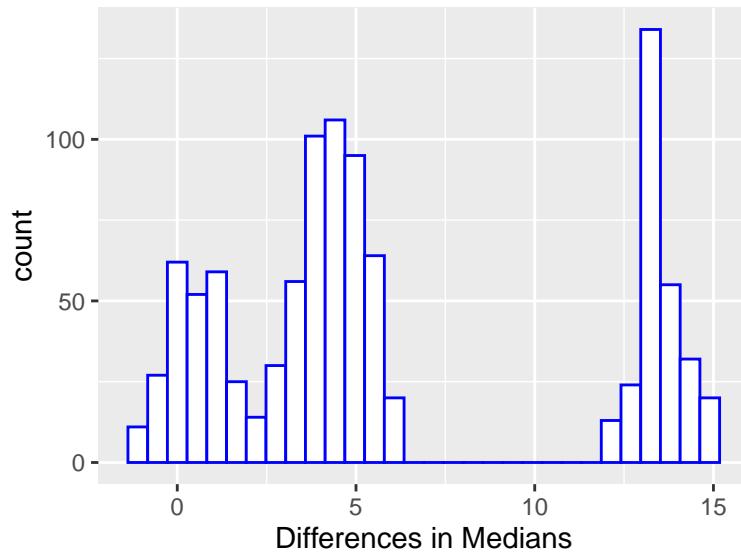
- We don't have a nonparametric interval for this
- Instead get Hodges-Lehman estimate
- Assume data come from distributions with same shape and differ only in location
- Considers all possible differences between sample 1 and sample 2 using male-female data on P. 7-13
- An estimate of the median difference (males - females) is the median of these 12 differences, with the 3rd and 10th largest values giving an (approximate) 0.95 CI
- Median estimate = 4.5, 0.95 CI = [1, 13]
- Specific formulas found in Altman, pages 40-41
- Bootstrap
 - General method, not just for medians
 - Non-parametric, does not assume symmetry (but may not be accurate)
 - Iterative method that repeatedly samples from the original data
 - Algorithm for creating a 0.95 CI for the difference in two medians
 1. Sample *with replacement* from sample 1 and sample 2
 2. Calculate the difference in medians, save result
 3. Repeat Steps 1 and 2 1000 times
 - A (naive) 0.95 CI is given by the 25th and 975th largest values of your 1000 median differences
 - For the male/female data, median estimate = 4.5, 0.95 CI = [-0.5, 14.5], which agrees with the conclusion from a WMW rank sum test ($p = 0.15$). Note that the more accurate CI for the Hodges-Lehman estimate of [-1, 15] was given earlier (output of `wilcox.test`).

```
diffs <- numeric(1000)
set.seed(13)
for(i in 1 : 1000) diffs[i] <-
  median(sample(male, replace=TRUE)) - median(sample(female, replace=TRUE))
```

```
ggplot(data.frame(diffs), aes(x=diffs)) + xlab('Differences in Medians') +  
  geom_histogram(bin_width=.01, color='blue', fill='white')
```

```
quantile(diffs, c(0.025, 0.975))
```

```
2.5% 97.5%  
-0.5 14.5
```



But recall that the Wilcoxon test does not really test the difference in medians but rather the median of all differences.

7.5

Strategy

- Don't assess normality of data
- Use nonparametric test in any case, to get P -values
- Use nonparametric confidence intervals for means and medians^c which will be more in conformance to what the nonpar. test is testing
- To obtain nonparametric confidence limits for means and differences in means, the bootstrap percentile method may easily be used and it does not assume symmetry of the data distribution

^cA good nonparametric confidence for a population mean that does not even assume a symmetric distribution can be obtained from the bootstrap simulation procedure.

7.6

Generalization of the Wilcoxon/Kruskal-Wallis Test



- Proportional odds (PO) ordinal logistic model
- Contains Wilcoxon 2-sample and Kruskal-Wallis tests as special cases
 - numerator of the score test for the PO model, when there is only the grouping variable in the model, is exactly the Wilcoxon statistic
- Special case of PO model is the ordinary binary logistic model
- Advantages over nonparametric tests:
 - can adjust for covariates
 - more accurate P -values even with extreme number of tied values
 - provides a framework for consistent pairwise comparisons^d
 - provides estimates of means, quantiles, and exceedance probabilities
 - sets the stage for a Bayesian PO model, so can get a Bayesian Wilcoxon test
- Other ordinal response models are available, e.g., Cox proportional hazards model
- These models are *semiparametric models*
 - parametric in additivity and linearity (by default) assumptions
 - nonparametric in not assuming a distribution for the response variable
- Like nonparametric tests, P -values are unaffected by monotonic transformations of Y

^dWhen using the Kruskal-Wallis test followed by pairwise Wilcoxon tests, these pairwise tests can be inconsistent with each other, because they re-rank the data based only on two groups, destroying the transitivity property, e.g. treatment A can be better than B which is better than C but C is better than A.

- If the response variable Y has k distinct values y_1, y_2, \dots, y_k in the sample, semi-parametric models have $k - 1$ intercepts
- Binary logistic model deals with prob. of only one event ($Y = 1$ vs. $Y = 0$)
- For ordinal Y there are $k - 1$ events
- Model these as cumulative probabilities to make use of ordering of Y values
- Model: $P(Y \geq y|X) = \frac{1}{1+\exp[-(\alpha_y + \beta_1 x_1 + \beta_2 x_2 + \dots)]}$
- α_y is the j^{th} intercept when $y = y_{j+1}$, e.g. the first intercept corresponds to the second lowest distinct Y value y_2
- Special case: 2-group problem: $P(Y \geq y|\text{group}) = \frac{1}{1+\exp[-(\alpha_y + \beta_1[\text{group B}])]}$
 - $\exp(\beta_1)$ is the ratio of odds that $Y \geq y$ in group B vs. $Y \geq y$ in group A, for all $y > y_1$
 - as before $[x]$ is the 0-1 indicator variable for x being true
 - $\beta_1 > 0 \rightarrow Y$ values higher in group B
 - $k = 2 \rightarrow$ model is the binary logistic model (where we take $\alpha_1 = \beta_0$)
- These intercepts $\alpha_1, \alpha_2, \dots, \alpha_{k-1}$ encode the entire empirical distribution of Y for one of the groups
 - → the model assumes nothing about the Y distribution
 - it only assumes how the distribution of Y for one type of subject is connected to the distribution for another type of subject
 - PO model for a two-group problem assumes that the logit of the two cumulative distribution functions are parallel
 - if PO doesn't hold, PO model may still be better than alternatives
 - PO is also what the Wilcoxon/Kruskal-Wallis test assumes to have optimal power

- don't need an α for the lowest observed Y value since $P(Y \geq \text{minimum } Y) = 1$
- R `rms` package `orm` function fits the PO model^e and is especially made for continuous Y , with fast run times for up to 6000 intercepts

7.6.1

Kruskal-Wallis Test

- Notice we haven't described rank ANOVA—the Kruskal-Wallis test
- Don't need it; just form a PO model with more than one indicator variable
- E.g., to test for any differences among four groups A B C D form 3 indicator variables for B C D and let A be the reference cell that corresponds to the α intercepts
 - model is $\text{logit}P(Y \geq y | \text{group}) = \alpha_y + \beta_1[B] + \beta_2[C] + \beta_3[D]$
- Use the likelihood ratio χ^2 test from this model to test the global null hypothesis A=B=C=D with 3 d.f.
- Solves the transitivity problem mentioned earlier
- Can obtain consistent pairwise comparisons by forming odds ratios for any comparison
 - e.g. C:A comparison will use $\exp(\hat{\beta}_2)$
 - C:B comparison OR: $\exp(\hat{\beta}_2 - \hat{\beta}_1)$
- As before can convert the ORs to differences in medians/means because unlike the original nonparametric tests, the PO model can be used to obtain many types of predictions^f
- Illustrate this by a non-PO example, checking to see how well the PO model can recover the sample means when assuming (the slightly incorrect) PO

^e`orm` also fits other models using link functions other than the logit.

^fThe predicted mean for a set of covariate settings is obtained by using all the intercepts and β s to get exceedance probabilities for $Y \geq y$, taking successive differences in those probabilities to get cell probabilities that $Y = y$, then multiplying cell probabilities by the y value attached to them, and summing. This is the formula for the mean for a discrete distribution.

- Take 4 samples from normal distributions with the same variances but different means
- Also show how to compare two of the samples without re-ranking the data as inconsistent Wilcoxon tests would do

```
set.seed(1)
group <- rep(c('A','B','C','D'), 100)
y <- rnorm(400, 100, 15) + 10*(group == 'B') + 20*(group=='C') + 30*(group=='D')
require(rms)
options(prType='latex')
dd <- datadist(group, y); options(datadist='dd')
f <- orm(y ~ group)
f    # use LR chi-square test as replacement for Kruskal-Wallis
```

Logistic (Proportional Odds) Ordinal Regression Model

`orm(formula = y ~ group)`

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	400	LR χ^2	193.31	R^2	0.383	ρ	0.633
Distinct Y	400	d.f.	3	g	1.532		
$Y_{0.5}$	115.0935		$\Pr(> \chi^2) <0.0001$	g_r	4.626		
max $ \frac{\partial \log L}{\partial \beta} $	2×10^{-6}	Score χ^2	193.21	$ \Pr(Y \geq Y_{0.5}) - \frac{1}{2} $	0.256		
			$\Pr(> \chi^2) <0.0001$				

	$\hat{\beta}$	S.E.	Wald Z	$\Pr(> Z)$
group=B	1.4221	0.2579	5.51	<0.0001
group=C	2.6624	0.2762	9.64	<0.0001
group=D	3.6606	0.2925	12.52	<0.0001

```
# Derive R function to use all intercepts and betas to compute predicted means
M <- Mean(f)
Predict(f, group, fun=M)
```

group	yhat	lower	upper
A	99.32328	95.87128	102.8162
B	111.21326	108.05575	114.3752
C	121.63880	118.56543	124.6699
D	129.70290	126.48067	132.8164

Response variable (y):

Limits are 0.95 confidence limits

```
# Compare with sample means
summarize(y, group, smean.cl.normal)
```

group	y	Lower	Upper
1 A	98.72953	95.81508	101.6440
2 B	111.69464	108.61130	114.7780
3 C	121.80841	118.93036	124.6865
4 D	130.05275	127.40318	132.7023

```
# Compare B and C
k ← contrast(f, list(group='C'), list(group='B'))
k
```

Contrast	S.E.	Lower	Upper	Z	Pr(> z)
11	1.240366	0.2564632	0.7377076	1.743025	4.84 0

Confidence intervals are 0.95 individual intervals

```
# Show odds ratios instead of differences in betas
print(k, fun=exp)
```

Contrast	S.E.	Lower	Upper	Z	Pr(> z)
11	3.456879	NA	2.091136	5.714604	4.84 0

Confidence intervals are 0.95 individual intervals

7.6.2

PO Re-analysis

- Reconsider the calprotectin data analyzed in Section 7.3.1
- Wilcoxon: $P = 0.0068, c = 0.837$
- Frequentist PO model:

```
require(rms)
```

```
options(prType='latex')
dd ← datadist(calpro, endo); options(datadist='dd')
f ← orm(calpro ~ endo)
print(f, intercepts=TRUE)
```

Logistic (Proportional Odds) Ordinal Regression Model

```
orm(formula = calpro ~ endo)
```

Frequencies of Responses

18	30	38	57	61	86	114	168	244	392	483	627	726	781	910	925
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1027	1226	2500													
			1	1	8										

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	26	LR χ^2	9.84	R^2	0.317	ρ	0.547
Distinct Y	19	d.f.	1	g	1.222		
$Y_{0.5}$	726	$\Pr(> \chi^2)$	0.0017	g_r	3.395		
$\max \frac{\partial \log L}{\partial \beta} $	5×10^{-5}	Score χ^2	9.86	$ \Pr(Y \geq Y_{0.5}) - \frac{1}{2} $	0.251		
		$\Pr(> \chi^2)$	0.0017				

	$\hat{\beta}$	S.E.	Wald Z	$\Pr(> Z)$
$y \geq 30$	2.0969	1.0756	1.95	0.0512
$y \geq 38$	1.3395	0.8160	1.64	0.1007
$y \geq 57$	0.8678	0.7135	1.22	0.2239
$y \geq 61$	0.4733	0.6689	0.71	0.4792
$y \geq 86$	0.1122	0.6575	0.17	0.8645
$y \geq 114$	-0.1956	0.6558	-0.30	0.7655
$y \geq 168$	-0.4710	0.6608	-0.71	0.4760
$y \geq 244$	-0.7653	0.6868	-1.11	0.2652
$y \geq 392$	-1.0953	0.7427	-1.47	0.1403
$y \geq 483$	-1.4155	0.8015	-1.77	0.0774
$y \geq 627$	-1.6849	0.8383	-2.01	0.0445
$y \geq 726$	-1.9227	0.8641	-2.23	0.0261
$y \geq 781$	-2.1399	0.8836	-2.42	0.0154
$y \geq 910$	-2.3439	0.8993	-2.61	0.0092
$y \geq 925$	-2.5396	0.9128	-2.78	0.0054
$y \geq 1027$	-2.7312	0.9249	-2.95	0.0031
$y \geq 1226$	-2.9224	0.9365	-3.12	0.0018
$y \geq 2500$	-3.1166	0.9482	-3.29	0.0010
endo=Moderate or Severe Activity	2.7586	0.9576	2.88	0.0040

- Intercept -3.1166 corresponds Y being at or above the upper detection limit
- Use the likelihood ratio (LR) χ^2 test from the model
- To estimate an exceedance probability just select the corresponding intercept and compute as for a binary logistic model
- The 18 intercepts for 19 distinct Y values represent the logit of the empirical cumulative distribution function for the no/mild reference group if the two groups

are in proportional odds^g. Add 2.7586 to those intercepts to get the logit CDF for the moderate/severe group.

- Compute odds ratio and CI

```
summary(f, endo='No or Mild Activity')
```

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
endo — Moderate or Severe Activity:No or Mild Activity	1	2		2.7586	0.95757	0.88175	4.6354
Odds Ratio	1	2		15.7770		2.41510	103.0700

- The above odds ratio of 15.8 is the odds of having calprotectin $\geq y$ in the moderate/severe activity group vs. the no/mild activity group
 - By the PO assumption this odds ratio is the same for all y

- Simulations provided an empirical conversion of the PO regression coefficient to c :

```
b <- coef(f) ['endo=Moderate or Severe Activity']
cindex <- plogis((b - 0.0029) / 1.5405)
cindex
```

```
endo=Moderate or Severe Activity
0.8567812
```

Compare this to the exact value of 0.837.

- From the fitted PO model obtain for each group, compute along with sample estimates:
 - prob. calprotectin at or above the upper limit of normal
 - mean
 - median

- In the output of Predict() see the point estimates under $yhat$, starting with the estimates for $P(Y \geq 2500)$, i.e., marker value at or above the upper detection limit

```
ex <- ExProb(f)
exceed <- function(lp) ex(lp, y=2500)
ymean <- Mean(f)
yquant <- Quantile(f)
ymed <- function(lp) yquant(0.5, lp=lp)
Predict(f, endo, fun=exceed)
```

^gThe intercepts really represent the logit of one minus the CDF, moved one Y value.

	endo	yhat	lower	upper
1	No or Mild Activity	0.04242948	0.008080776	0.1941982
2	Moderate or Severe Activity	0.41144485	0.209594428	0.6482557

Response variable (y):

Limits are 0.95 confidence limits

```
# Compute empirical exceedance probabilities
tapply(calpro >= 2500, endo, mean)
```

No or Mild Activity	Moderate or Severe Activity
0.1250000	0.3888889

```
# Note that imposing PO assumption made modeled means closer together than
# stratified sample means
Predict(f, endo, fun=ymean)
```

	endo	yhat	lower	upper
1	No or Mild Activity	300.259	91.55091	851.9429
2	Moderate or Severe Activity	1387.660	895.58358	1868.2181

Response variable (y):

Limits are 0.95 confidence limits

```
tapply(calpro, endo, mean)
```

No or Mild Activity	Moderate or Severe Activity
400.000	1372.944

```
Predict(f, endo, fun=ymed)
```

	endo	yhat	lower	upper
1	No or Mild Activity	69.59518	23.59636	488.0126
2	Moderate or Severe Activity	940.32171	549.13891	1653.9661

Response variable (y):

Limits are 0.95 confidence limits

```
tapply(calpro, endo, median)
```

No or Mild Activity	Moderate or Severe Activity
87.5	976.0

- Note: confidence intervals for these derived quantities are approximate

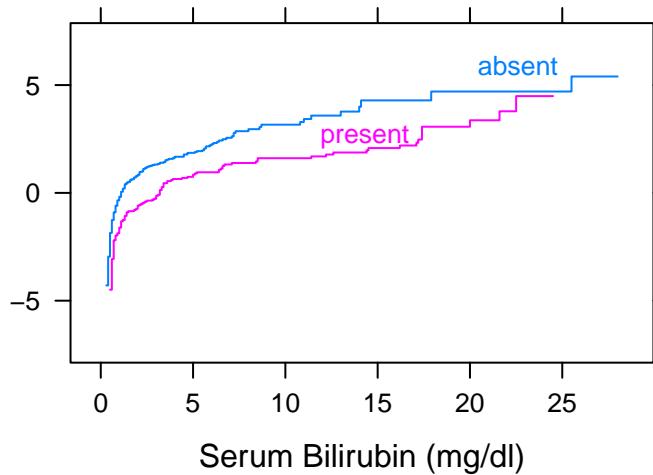
7.7

Checking Assumptions of the Wilcoxon Test

- Proportional odds (PO) model and its special case the Wilcoxon test assume PO
- What does it mean to *assume* PO?
 - Under H_0 : the two distributions are **identical** there is no assumption, i.e., type I error probability will behave as advertised
 - Under H_1 the test may still work OK but it will not be *optimal* unless PO holds
- To check PO:
 - Compute the empirical cumulative distribution function (ECDF) for the response variable, stratified by group (see Section 4.3.3)
 - Take the logit transformation of each ECDF
 - Check for parallelism
 - Linearity would be required **only** if using a parametric logistic distribution instead of using our semiparametric PO model
- Parametric *t*-test requires parallelism **and** linearity when the ECDFs are normal-inverse transformed
 - linearity: normal distribution (like q-q plot)
 - parallelism: equal variances
- Problem with assumption checking is that with small samples ECDFs are too noisy to see patterns clearly
- Example from a larger dataset: Mayo Clinic Primary Biliary Cirrhosis Dataset
- Compare distribution of serum bilirubin for those patients with spider veins vs. those without

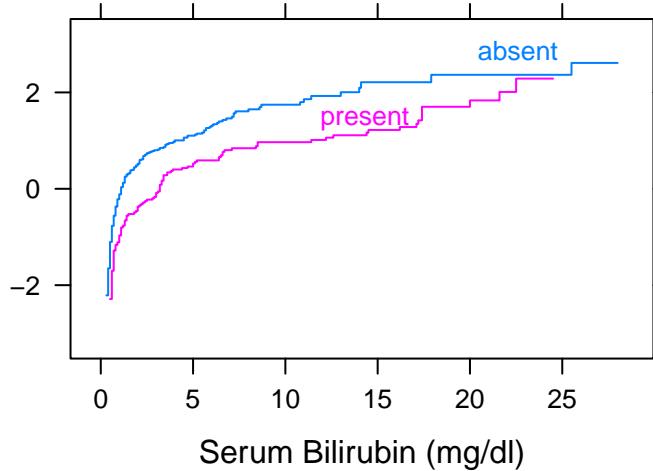
```
getHdata(pbc)
# Take logit of ECDF
```

```
Ecdf(~ bili, group = spiders, data=pbc, fun=qlogis)
```



- The curves are primarily parallel (even at the far left, despite the optical illusion)
- Nonlinearity is irrelevant
- Check *t*-test assumptions

```
Ecdf(~ bili, group=spiders, data=pbc, fun=qnorm)
```



- Curves are primarily parallel (variances are equal)
- **But** they are not straight lines as required by *t*-test normality assumption

7.8

Power and Sample Size

7.8.1

Normal Approximation

- Common to get power/sample size estimates for the Wilcoxon two-sample comparison using the unpaired t -test power formula
- Are assuming normality and (usually) equal variances
- To reflect the slight inefficiency of the Wilcoxon two-sample test if normality were to magically hold, multiply the t -test sample size by $\frac{\pi}{3} = 1.047$
- When the response within-group distribution is far from normal this approach is suspect
 - e.g., Y has many ties at one value, has a floor or ceiling effect, is asymmetric, or has heavy tails
- Need a general approach

7.8.2

More on Relative Efficiency

- Relative efficiency of $\frac{3}{\pi}$ for the Wilcoxon 2-sample test can be derived as a correlation coefficient
- As $n \rightarrow \infty$ it is the squared correlation between the weights Wilcoxon gives to order statistics (sorted data values) and the optimal weights
- Wilcoxon is a *linear rank statistic* with weights equal to ordinary ranks
- Optimal linear rank test (normal scores test) for a normal distribution uses the probit link (normal inverse weights), i.e., $\Phi^{-1}\left(\frac{\text{ranks}}{n+1}\right)$

- Compute correlation of ordinary ranks with normal scores

```
for(n in c(10, 100, 1000, 10000, 100000, 1000000)) {
  ranks  ← 1 : n
  zranks ← qnorm(ranks / (n + 1))
  cat('n:', n, ' r2:', cor(ranks, zranks)^2, '\n')
}
```

```
n: 10    r2: 0.9923288
n: 100   r2: 0.9688625
n: 1000  r2: 0.958053
n: 10000 r2: 0.9554628
n: 1e+05 r2: 0.955008
n: 1e+06 r2: 0.9549402
```

```
cat('3/pi: ', 3 / pi, '\n')
```

```
3/pi: 0.9549297
```

7.8.3

Tailoring Power Calculations to the Wilcoxon Test

- Whitehead¹¹² derived simple formulas related to the proportional odds model score test
- Formulas assume that a frequency-tabulated distribution estimate is available for the combined groups
- Power is computed as a function of the group 2 : group 1 odds ratio for exceedance probabilities
- See example below for conversion of ORs to differences in means or medians
 - OR=1 → distributions are the same, so differences in means/medians are zero
- See R `Hmisc` package `popower` and `posamsize` functions

7.8.4

Discrete Y

- Example: response variable has clumping at zero (with prob. 0.3) and is otherwise uniformly distributed over the values 1, 2, 4, 8, 16, 32, 64

- note: actual data values do not affect power Calculations
- don't come into play until translate to means/medians

```
p ← c(.3, rep(.1, 7))
popower(p, 1.25, 1000) # compute power to detect OR=1.25, combined N=1000
```

```
Power: 0.516
Efficiency of design compared with continuous response: 0.966
```

```
posamsize(p, 1.25, power=0.9) # N for power=0.9
```

```
Total sample size: 2621.4
Efficiency of design compared with continuous response: 0.966
```

- Show how cell probabilities are translated by OR=1.25, and compute the mean and median of Y for a series of ORs for simpler interpretation

```
pomodm(p=p, odds.ratio=1.25)
```

```
[1] 0.25531915 0.09250694 0.09661836 0.10101010 0.10570825 0.11074197 0.11614402
[8] 0.12195122
```

```
x ← c(0, 2^(0:6))
sum(p * x) # check mean with OR=1
```

```
[1] 12.7
```

```
ors ← c(1, 1.05, 1.1, 1.2, 1.25, 1.5, 2)
w ← matrix(NA, nrow=length(ors), ncol=2,
           dimnames=list(OR=ors, c('mean', 'median')))
i ← 0
for(or in ors) {
  i ← i + 1
  w[i, ] ← pomodm(x, p, odds.ratio=or)
}
w
```

OR	mean	median
1	12.70000	3.000000
1.05	13.14602	3.364286
1.1	13.58143	3.709091
1.2	14.42238	4.350000
1.25	14.82881	4.650000
1.5	16.73559	6.000000
2	20.03640	9.900000

7.8.5

Gaussian Y

- Suppose response variable for control group has a normal distribution with mean 100 and SD 10
- Start by assuming the experimental arm has the same distribution as control except with the mean shifted upwards 3 units
- This will result in non-proportional odds so the Wilcoxon test is not optimal but will still be 0.95 efficient
- When the sample size per group is 150, the power of the t -test to detect a 3-unit difference in means is:

```
require(pwr)
pwr.t.test(d=3 / 10, n=150, sig.level=0.05, type='two.sample')
```

Two-sample t test power calculation

```
n = 150
d = 0.3
sig.level = 0.05
power = 0.7355674
alternative = two.sided
```

NOTE: n is number in *each* group

- To get the power of the Wilcoxon test when both populations have a normal distribution, we can easily use simulation

```
s <- 1000 # number of simulated trials
pval <- numeric(s)
set.seed(1) # so can reproduce results
for(i in 1 : s) {
  y1 <- rnorm(150, 100, 10)
  y2 <- rnorm(150, 103, 10)
  w <- wilcox.test(y1, y2)
  pval[i] <- w$p.value
}
mean(pval < 0.05) # proportion of simulations with p < 0.05
```

[1] 0.713

```
# Simulate the power by actually running the prop. odds model 300 times
simRegOrd(300, nsim=400, delta=3, sigma=10)$power # slower
```

```
[1] 0.71
```

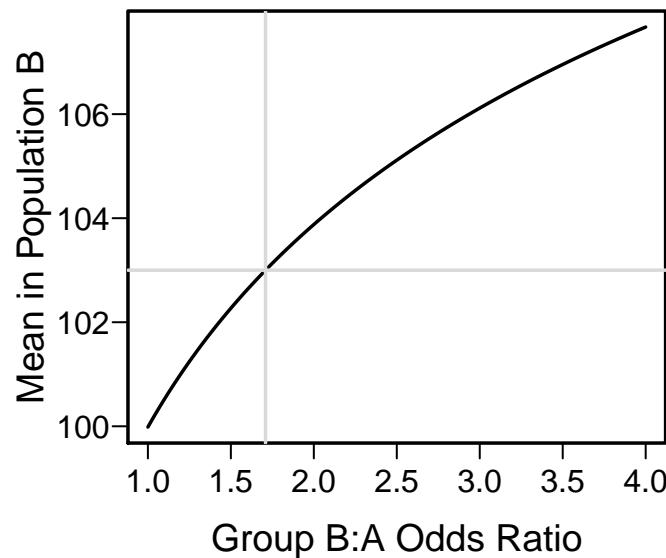
- For the Wilcoxon test to be optimal (PO holds) shifting the control distribution by an odds ratio will result in a non-Gaussian distribution for the experimental arm
- Solve for the odds ratio that shifts the mean from 100 to 103, assume PO and compute the power

```
# Use an arbitrary large sample to mimic population computations
m   ← 200000
y1  ← sort(rnorm(m, 100, 10))
ors ← means ← seq(1, 4, by=.025)
i   ← 0
for(or in ors) {
  i ← i + 1
  means[i] ← pomodm(y1, rep(1/m, m), odds.ratio=or)[‘mean’]
}
plot(ors, means, xlab='Group B:A Odds Ratio',
      ylab='Mean in Population B', type='l')
abline(h=103, col=gray(.85))
needed.or ← approx(means, ors, xout=103)$y
needed.or
```

```
[1] 1.708958
```

```
abline(v=needed.or, col=gray(.85))
# Compute power at that odds ratio assuming no ties in data
popower(rep(1/300, 300), odds.ratio=needed.or, n=300)
```

```
Power: 0.761
Efficiency of design compared with continuous response: 1
```

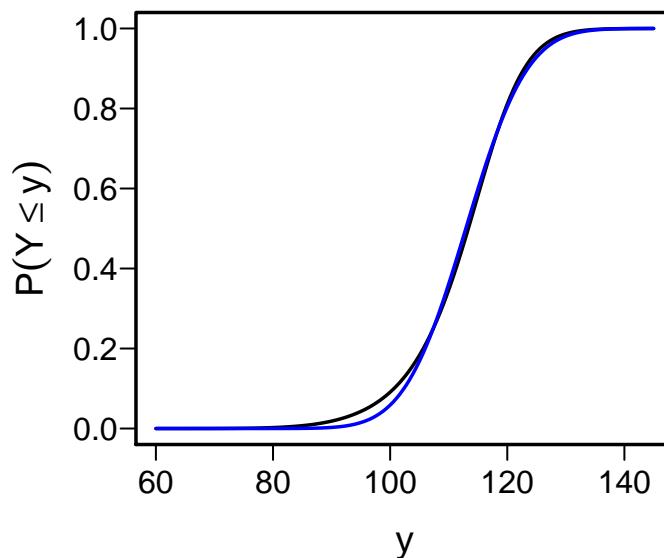


- Check how non-normal the experimental arm responses would be if PO holds and OR=10

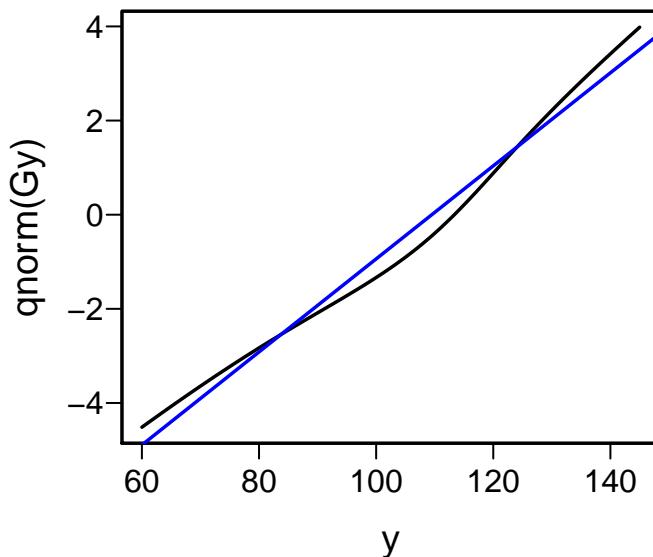
```
# First do this theoretically
# Control arm has Gaussian Y with mean 100, SD 10
# Create experimental arm distribution using OR=10
y ← seq(60, 145, length=150)
Fy ← 1 - pnorm(y, mean=100, sd=10)      # P(Y ≥ y | group A)
Gy ← 1 - plogis(qlogis(Fy) + log(10)) # P(Y ≥ y | group B)
# Plot new CDF vs. normal approximation agreeing at quartiles
plot(y, Gy, type='l', ylab=expression(P(Y ≤ y)))
qu ← approx(Gy, y, xout=c(0.25, 0.5, 0.75))$y
qu      # Q1, median, Q2
```

```
[1] 107.3628 113.3506 118.4894
```

```
s ← (qu[3] - qu[1]) / (qnorm(0.75) - qnorm(0.25))
mu ← qu[1] - s * qnorm(0.25)
lines(y, pnorm(y, mean=mu, sd=s), col='blue')    # Gaussian fit
```



```
# Theoretical q-q plot: check linearity of inverse normally transformed
# experimental arm distribution
plot(y, qnorm(Gy), type='l')
abline(lsfit(y, qnorm(Gy)), col='blue')
```



```
# Compute a new discrete distribution if we convert the control
# distribution using proportional odds
# Done by using a discrete distribution with 200,000 points
p <- pomodm(p=rep(1/m, m), odds.ratio=10)
range(p) # control arm: all 1/200000
```

```
[1] 5.000023e-07 4.999775e-05
```

```
wtd.mean(y1, p) # mean shifted by about 12 units
```

```
[1] 112.4126
```

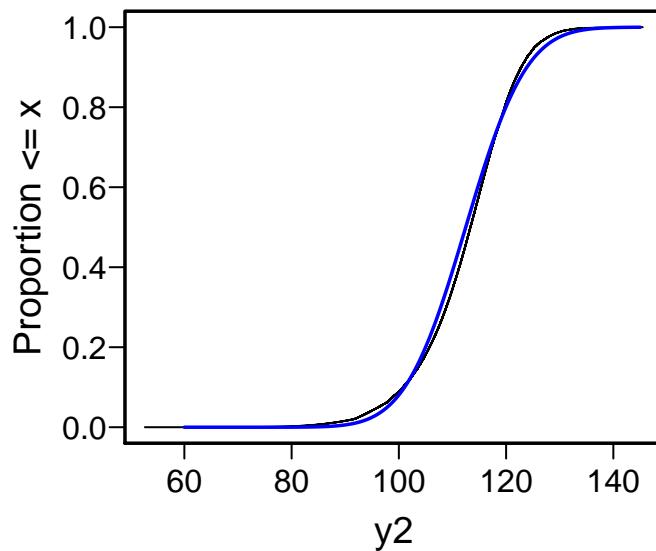
```
# Form new distribution by repeating each observation a number
# of times equal to the ratio of the new probability to the
# minimum of all new probabilities
y2 <- rep(y1, round(p / min(p))) # 2M obs instead of 200K
mean(y2)
```

```
[1] 112.4714
```

```
quantile(y2, c(.25, .5, .75))
```

	25%	50%	75%
107.4240	113.3446	118.5467	

```
# The following plot is similar to the previous one
Ecdf(y2, subtitles=FALSE)
lines(y, pnorm(y, mean=mean(y2), sd=sd(y2)), col='blue')
```



- Little non-normality of the 2nd group if the treatment effect operates by multiplying the odds that $Y \geq y$ instead of incrementing the mean
 - relates to similarity of normal and logistic distributions

7.8.6

Heavy-tailed Y

- Get power to detect a shift in mean of 0.3 units for a heavy-tailed control distribution (t with 4 d.f.) with 150 subjects per group
- Loss of efficiency of t -test
 - mean and SD are no longer optimal data summaries
- Can use above method to compute Wilcoxon power quickly if willing to assume PO
- Let's not assume PO, and instead use simulation
- Compare with power of t -test
- Do for both null and non-null cases to verify type I error prob.

```
s <- 1000      # number of simulated trials
pvalt <- pvalw <- numeric(s)
set.seed(1)    # so can reproduce results
for(delta in c(0, 0.3)) {
```

```
for(i in 1 : s) {  
  y1 ← rt(150, 4)  
  y2 ← rt(150, 4) + delta  
  pvalt[i] ← t.test(y1, y2)$p.value  
  pvalw[i] ← wilcox.test(y1, y2)$p.value  
}  
cat('Delta:', delta, '\n')  
P ← function(x) round(mean(x), 2)  
cat('Proportion of simulations with W p-value < t p-value:',  
    P(pvalw < pvalt), '\n')  
cat('Mean p-value for t:', P(pvalt), '\n')  
cat('Mean p-value for W:', P(pvalw), '\n')  
cat('Power for t:', P(pvalt < 0.05), '\n')  
cat('Power for W:', P(pvalw < 0.05), '\n\n')
```

```
Delta: 0  
Proportion of simulations with W p-value < t p-value: 0.51  
Mean p-value for t: 0.5  
Mean p-value for W: 0.49  
Power for t: 0.05  
Power for W: 0.06  
  
Delta: 0.3  
Proportion of simulations with W p-value < t p-value: 0.73  
Mean p-value for t: 0.17  
Mean p-value for W: 0.12  
Power for t: 0.47  
Power for W: 0.6
```

- Hmisc simRegOrd function can also simulate power for an adjusted two-sample comparison if there is one adjustment covariate

7.9

Bayesian Proportional Odds Model

- PO model and other cumulative probability semiparametric ordinal regression models are readily extended to a Bayesian framework
- Need special care in selecting priors for the intercepts for the continuous Y case
- Nathan James of Vanderbilt University has an implementation using Stan available at
github.com/ntjames/bayes_cpm
- See also the R `brms` package: bit.ly/brms-ordinal and this discussion:
github.com/paul-buerkner/brms/issues/762
- Bayesian version of the Wilcoxon test is the posterior probability that $\beta_1 > 0$ in the PO model
- Advantages of Bayes for PO models:
 - does not need approximations such as large sample normality of $\hat{\beta}$ or χ^2 distribution approximation to likelihood ratio test statistic
 - inference is more interpretable and direct
 - can bring outside information to the analysis
 - can incorporate shrinkage/penalization/skepticism and still have exact inference
 - automatically obtain exact distributions and credible intervals for derived quantities^h, e.g. mean, quantiles, differences in means and quantiles, differences in exceedance probs, $P(Y = y|X)$
 - can relax PO assumption without huge instabilities that result from using polytomous logistic models; prior distributions can favor PO while allowing non-PO

^hOne merely takes each posterior sample for the α s and β s and computes the quantity of interest, thereby automatically generating posterior samples for the derived quantity for which quantiles can compute credible intervals, etc.

Chapter 8

Correlation



8.1

Overview

Outcome	Predictor	Normality?	Linearity?	Analysis Method
Interval	Binary	Yes		2-sample <i>t</i> -test or linear regression
Ordinal	Binary	No		Wilcoxon 2-sample test
Categorical	Categorical			Pearson χ^2 test
Interval	Interval	Yes	Yes	Correlation or linear regression
Ordinal	Ordinal	No	No	Spearman's rank correlation

- Examine association between continuous/interval outcome (y) and continuous/interval predictor (x)
- Scatterplot of y versus x

8.2

Pearson's correlation coefficient

$$\bullet r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- Range: $-1 \leq r \leq 1$
- Correlation coefficient is a unitless index of strength of association between two variables (+ = positive association, - = negative, 0 = no association)
- Measures the linear relationship between X and Y
- Can test for significant association by testing whether the population correlation is zero

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

which is identical to the t -test used to test whether the population r is zero; d.f.= $n - 2$.

- Use probability calculator for t distribution to get P -value (2-tailed if interested in association in either direction)
- 1-tailed test for a positive correlation between X and Y tests H_0 : when $X \uparrow$ does $Y \uparrow$ in the population?
- Confidence intervals for population r calculated using Fisher's Z transformation

$$Z = \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right)$$

- For large n , Z follows a Normal distribution with standard error $\frac{1}{\sqrt{n-3}}$
- To calculate a confidence interval for r , first find the confidence interval for Z then transform back to the r scale

$$\begin{aligned} Z &= \frac{1}{2} \log_e \left(\frac{1+r}{1-r} \right) \\ 2 * Z &= \log_e \left(\frac{1+r}{1-r} \right) \end{aligned}$$

$$\begin{aligned}
 \exp(2 * Z) &= \left(\frac{1+r}{1-r} \right) \\
 \exp(2 * Z) * (1-r) &= 1+r \\
 \exp(2 * Z) - r * \exp(2 * Z) &= 1+r \\
 \exp(2 * Z) - 1 &= r * \exp(2 * Z) + r \\
 \exp(2 * Z) - 1 &= r (\exp(2 * Z) + 1) \\
 \frac{\exp(2 * Z) - 1}{\exp(2 * Z) + 1} &= r
 \end{aligned}$$

- Example (Altman 89-90): Pearson's r for a study investigating the association of basal metabolic rate with total energy expenditure was calculated to be 0.7283 in a study of 13 women. Derive a 0.95 confidence interval for r .

$$Z = \frac{1}{2} \log_e \left(\frac{1+0.7283}{1-0.7283} \right) = 0.9251$$

The lower limit of a 0.95 CI for Z is given by

$$0.9251 - 1.96 * \frac{1}{\sqrt{13-3}} = 0.3053$$

and the upper limit is

$$0.9251 + 1.96 * \frac{1}{\sqrt{13-3}} = 1.545$$

A 0.95 CI for the population correlation coefficient is given by transforming these limits from the Z scale back to the r scale

$$\frac{\exp(2 * 0.3053) - 1}{\exp(2 * 0.3053) + 1} \text{ to } \frac{\exp(2 * 1.545) - 1}{\exp(2 * 1.545) + 1}$$

Which gives a 0.95 CI from 0.30 to 0.91 for the population correlation

```

n ← 13
r ← 0.7283
z.transform ← 0.5 * log((1 + r) / (1 - r))
clz ← z.transform + c(-1, 1) * qnorm(0.975) / sqrt(n - 3)
clr ← (exp(2 * clz) - 1) / (exp(2 * clz) + 1)
round(c(z.transform, clz, clr), 4)

```

```
[1] 0.9251 0.3053 1.5449 0.2962 0.9129
```

8.3

Spearman's Rank Correlation

- Pearson's r assumes linear relationship between X and Y
- Spearman's ρ (sometimes labeled r_s) assumes monotonic relationship between X and Y
 - when $X \uparrow$, Y always \uparrow or stays flat, or Y always \downarrow or stays flat
 - does not assume linearity
- $\rho = r$ once replace column of X s by their ranks and column of Y s by ranks
- To test $H_0 : \rho = 0$ without assuming linearity or normality, being damaged by outliers, or sacrificing much power (even if data are normal), use a t statistic:

$$t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

which is identical to the t -test used to test whether the population r is zero; d.f.= $n - 2$.

- Use probability calculator for t distribution to get P -value (2-tailed if interested in association in either direction)
- 1-tailed test for a positive correlation between X and Y tests H_0 : when $X \uparrow$ does $Y \uparrow$ in the population?

8.4

Correlation Examples

- Correlation difficult to judge by eye
- Example plots on following pages

```
# Generate 50 data points with Population correlations of 0, .2, .4, .6,
# .8, and .9 and plot results
require(ggplot2)
n <- 50
set.seed(123)
x <- rnorm(n, 5, 1)
d <- expand.grid(x=x, R=c(0, .2, .4, .6, .8, .9))
d <- transform(d, y = x + rnorm(nrow(d), 0,
                                 ifelse(R == 0, 5, sqrt(R ^ -2 - 1))))
sfun <- function(i) {
  x <- d$x[i]; y <- d$y[i]; R <- d$R[i][1]
  r <- cor(x, y)
  tr <- r * sqrt(n - 2) / sqrt(1 - r^2)
  rho <- cor(rank(x), rank(y))
  trho <- rho * sqrt(n - 2) / sqrt(1 - rho^2)
  label <- paste('True r:', R[1], ' r:', round(r,2), ' t:', round(tr,2),
                ' rho:', round(rho,2), ' t:', round(trho,2), sep=' ')
  names(label) <- R
  label
}
stats <- tapply(1 : nrow(d), d$R, sfun)
d$stats <- factor(stats[as.character(d$R)], unique(stats))

ggplot(d, aes(x=x, y=y)) + geom_point() + facet_wrap(~ stats) +
  theme(strip.text.x = element_text(size=7)) # Fig. 8.1
```

```
# Different scenarios that can lead to a correlation of 0.7

set.seed(123) # Fig. 8.2
rho <- 0.7; n <- 50
var.eps <- rho^2 - 1
x <- rnorm(n, 5, 1)
y <- x + rnorm(n, 0, sqrt(var.eps))
cor(x,y)
```

```
[1] 0.6951673
```

```
plot(x,y,xlab=' ',ylab=' ')
x <- c(1:20,30)
y <- c(1:20,6.2)
cor(x,y)
```

```
[1] 0.6988119
```

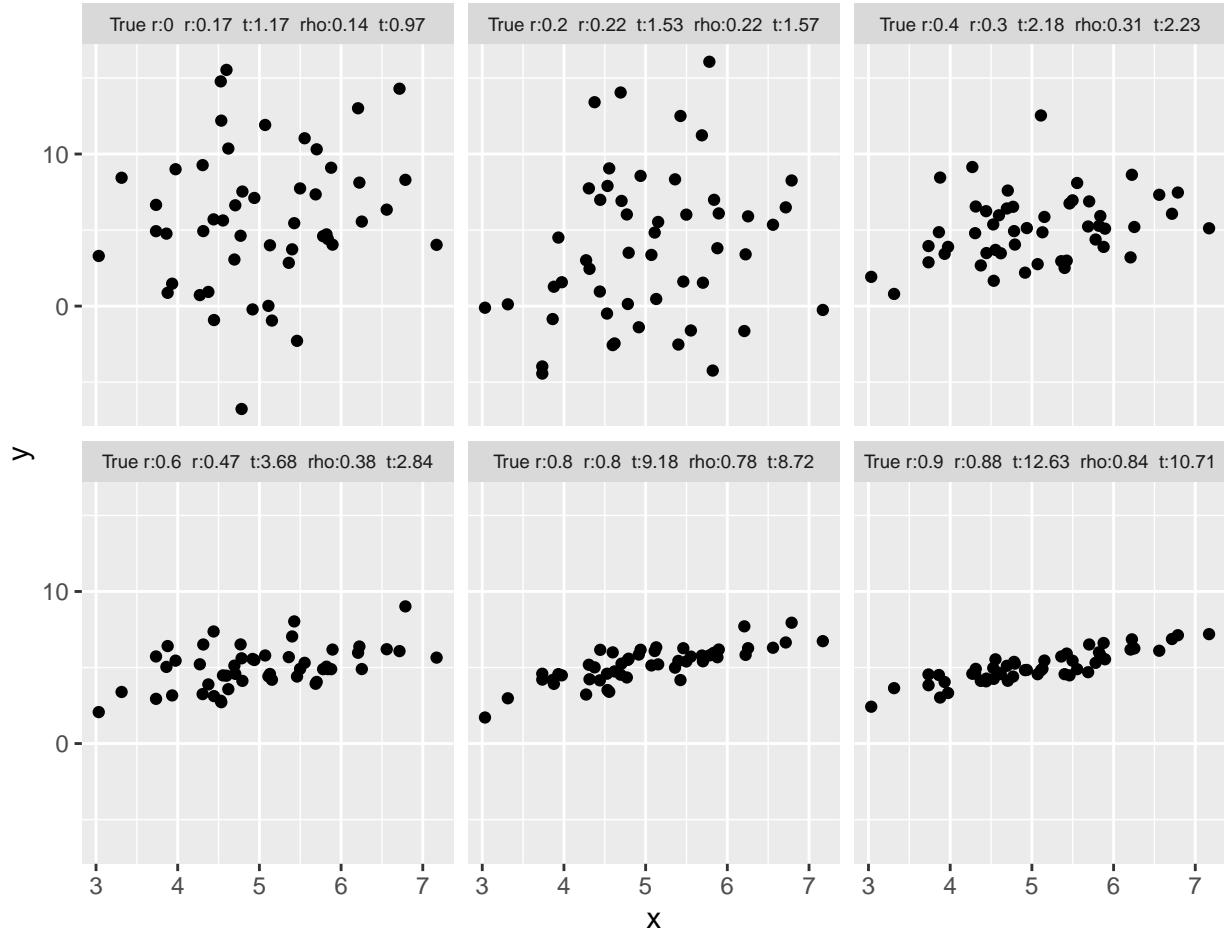


Figure 8.1: Samples of size $n = 50$ for X and Y are drawn from bivariate normal populations with true correlations ranging from 0.0 to 0.9. Pearson and Spearman sample correlations are shown for samples of size 50. Besides the population correlation coefficient, each panel is labeled with the estimated Pearson r , its t statistic, the estimated Spearman ρ , and its t statistic

```
plot(x,y,xlab=' ',ylab=' ')
set.seed(123)
x <- rnorm(40)
y <- rnorm(40)
x[21] <- y[21] <- 8.5
cor(x,y)
```

```
[1] 0.7014825
```

```
plot(x,y,xlab=' ',ylab=' ')
x <- rep(0:19,2)
y <- c(rep(.62,20),rep(2,20)) * x
cor(x,y)
```

```
[1] 0.701783
```

```
plot(x,y,xlab=' ',ylab=' ')
x <- -7:12
y <- x^2
cor(x,y)
```

```
[1] 0.6974104
```

```
plot(x,y,xlab=' ',ylab=' ')
set.seed(123)
tmp <- 1:20 / 2
x <- c(rnorm(20, tmp, 1), tmp + rnorm(20,14.5,1))
y <- c(rnorm(20, -tmp, 1), -tmp + rnorm(20,14.5,1))
cor(x,y)
```

```
[1] 0.703308
```

```
plot(x,y,xlab=' ',ylab=' ')
```

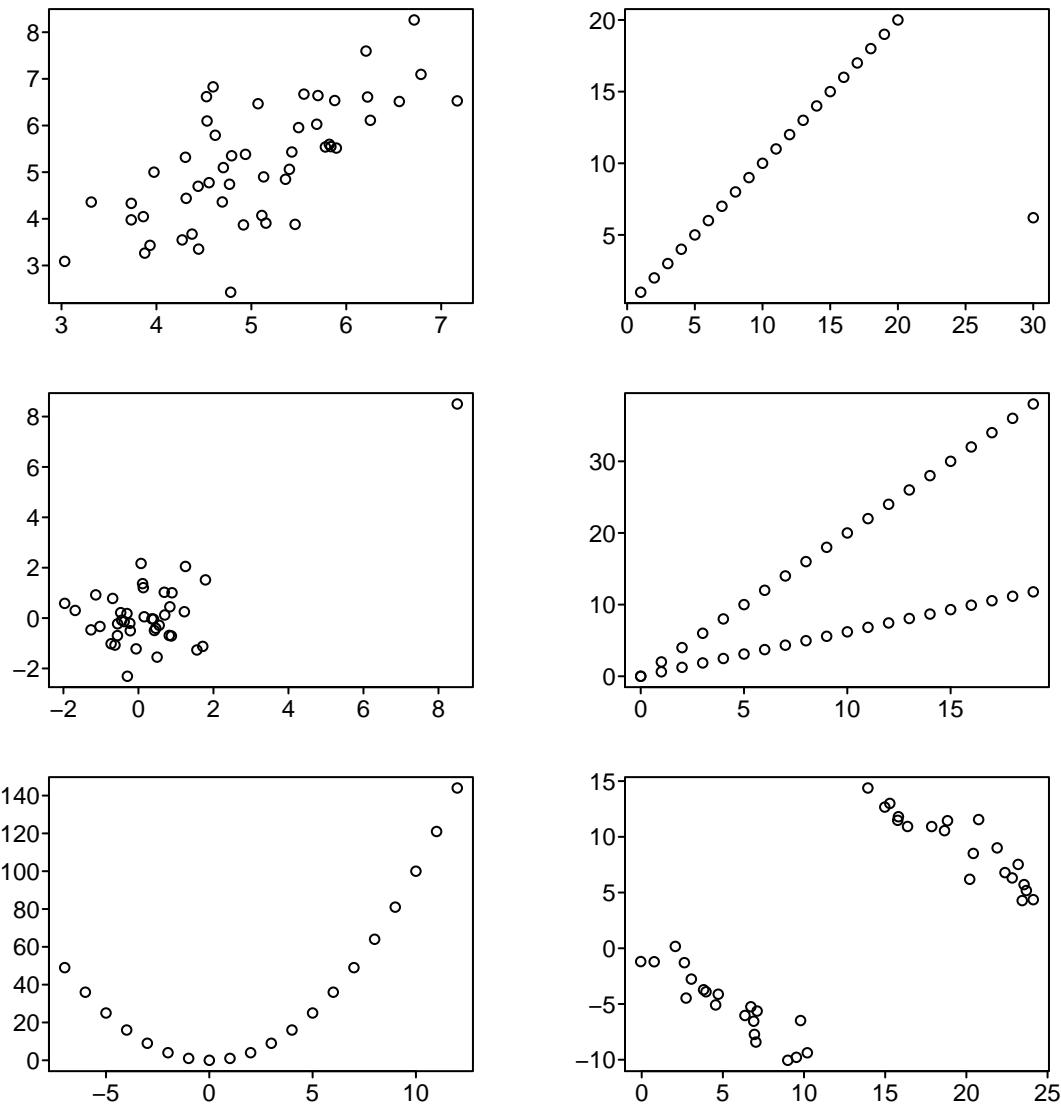


Figure 8.2: Different observed datasets that have the same correlation. All six plots have a sample Pearson correlation of 0.7.

8.5

Correlation and Agreement

- Compare two methods of measuring the same underlying value
 - Lung function measured using a spirometer (expensive, accurate) or peak flow meter (cheap, less accurate)
 - Two devices (oropharyngeal and conventional) used to measured acidity (pH) in the esophagus as a marker of reflux
- Typical (incorrect) approach begins with scatterplot of one method vs. the other with a 1:1 line indicating perfect agreement
- See Figure 4.11
- Incorrect approach would report a high correlation ($r = 0.90$) and conclude good agreement
- Problems with the correlation approach
 1. r measures the degree of linear association between two variables, not the agreement. If, for example, the Sandhill consistently gave pH values that were 0.5 unit higher than the Restech, we could still have high correlation, but poor agreement between the two devices. We can have high correlation if the two devices lie closely to any line, not just a 1:1 line that indicates perfect agreement.
 2. A change in scale does not affect correlation, but does influence agreement. For example, if the Sandhill always registered 2 times larger than the Restech, we would have perfect correlation but the agreement would get progressively worse for larger values of pH.
 3. Correlation depends on the range of the data so that larger ranges lead to larger correlations. This can lead to vary strange interpretations
 4. Tests of significance (testing if $r = 0$) are irrelevant to the question at hand, but often reported to demonstrate a significant association. The two devices are measuring the same quantity, so it would be shocking if we did not observe a highly significant p -value. A $p < .0001$ is not impressive. A regression analysis with a highly significant slope would be similarly unimpressive.

	r	ρ
all data	0.90	0.73
avg pH ≤ 4	0.51	0.58
avg pH > 4	0.74	0.65

Table 8.1: Pearson (r) and Spearman (ρ) correlations for Restech and Sandhill pH data. The correlation calculated using all of the data is larger than the correlation calculated using a restricted range of the data. However, it would be difficult to claim that the overall agreement is better than both the agreement when pH is less than 4 and when pH is greater than 4.

5. Data can have high correlation, but poor agreement. There are many examples in the literature, but even in our analysis with $r = 0.90$, the correlation is high, but we will show that the agreement is not as good as the high correlation implies.

See Chapter 16 for simple approaches to assessing agreement and analyzing observer variability studies.

8.5.1

Bland-Altman Plots

- See Bland and Altman (1986, Lancet)
- Earlier: Tukey mean-difference plot
- Create plots of the difference in measurements on the y-axis versus the average value of the two devices on the x-axis
- If the two devices agree, the difference should be about zero
- The average of the two devices is our best estimate of the true, unknown (pH) value that is we are trying to measure
- Measurements will often vary in a systematic way over the range of measurement. By plotting the difference versus the average, we can visually determine if the difference changes over our estimate of the truth.
- Solid line indicated the mean, dashed lines are approximate 0.95 confidence intervals (assuming Normality)

But there is controversy about what should be on the x -axis of the plot. Krouwer⁵⁸ concluded that:

- When the two measures have nearly equal variability, i.e., when comparing two “field measurements”, the Bland-Altman approach is preferred
- When one measurement is a “reference standard” having much less variation than the field measurement, the reference standard and not the average of the two measurements should be on the x -axis

```
require(Hmisc)
getHdata(esoph)
esoph$diff <- with(esoph, orophar - conv)
ggplot(esoph, aes(x=(conv + orophar)/2, y=diff)) +    # Fig. 8.3
  stat_binhex(aes(alpha=..count.., color=Hmisc::cut2(..count.., g=20)),
               bins=80) +
  stat_smooth() +
  geom_hline(yintercept = mean(esoph$diff, na.rm=TRUE) +
    c(-1.96, 0, 1.96) * sd(esoph$diff, na.rm=TRUE),
    linetype=c(2,1,2), color='brown') +
  xlab('Average of Conventional and Oropharyngeal pH') +
  ylab('Oropharyngeal Minus Conventional pH') +
  guides(alpha=FALSE, fill=FALSE, color=guide_legend(title='Frequency'))
```

- We will also consider differences in the two measurements over the time of day
- The added smooth curve is called a locally weighted scatterplot smooth (loess)

```
getHdata(esoph2)
ggplot(esoph2, aes(x=time, y=diffpH)) +    # Fig. 8.4
  geom_point(pch='.') + stat_smooth() +
  geom_hline(yintercept = 0, col='gray60') +
  scale_x_continuous(breaks=seq(16, 38, by=4),
    labels=c("4 PM", "8 PM", "12 AM",
    "4 AM", "8AM", "12 PM"),
    limits=c(14, 14+24)) +
  ylab('Average of Oropharyngeal Minus Conventional pH') +
  xlab('Time of Day')
```

8.5.2

Sample Size for r

- Without knowledge of population variances, etc., r can be useful for planning studies

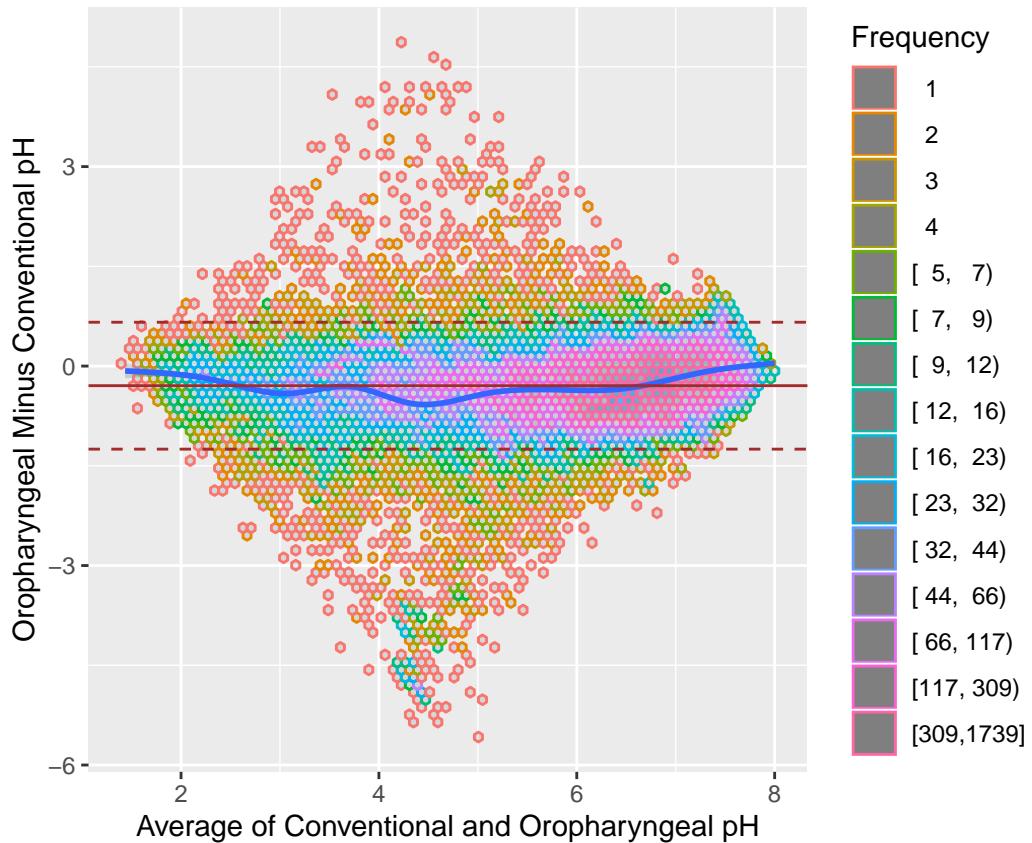


Figure 8.3: Bland-Altman plot for the oroesophageal and conventional pH measurements, using hexagonal binning because of the large sample size. The difference in pH measurements (oro. -conventional) is presented on the y -axis and the average of the two devices on the x -axis. We see poor agreement around pH values of 4-5

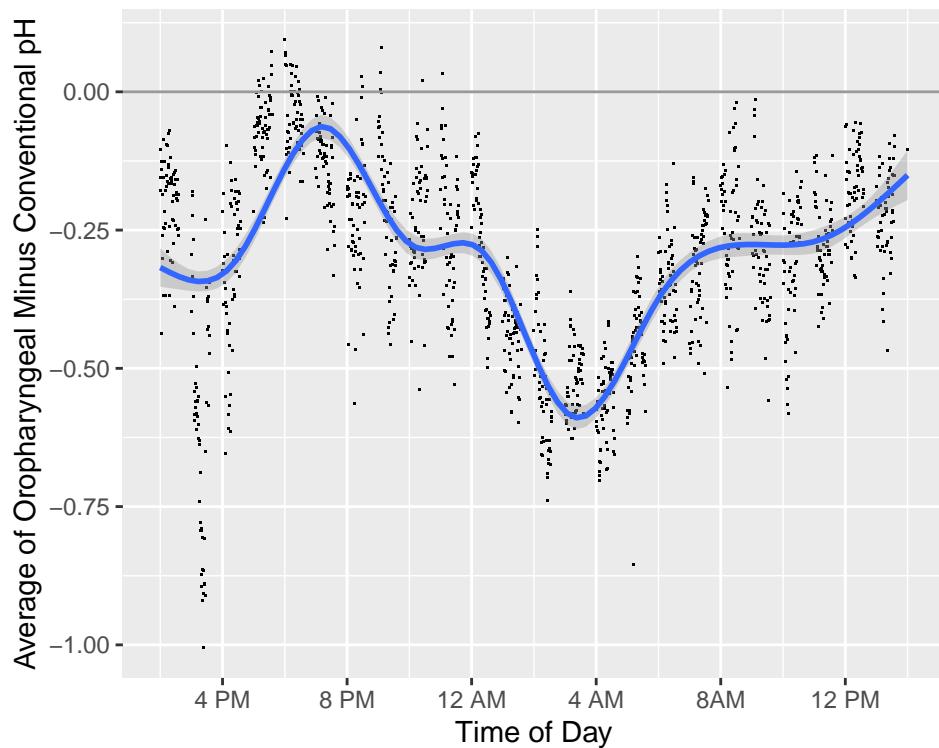


Figure 8.4: Difference in pH measurements (oro. - conventional) by time of day along with a loess smoother and pointwise 0.95 confidence bands. Is the difference modified by a subject being in a supine position rather than being upright?

- Choose n so that margin for error (half-width of C.L.) for r is acceptable
- Precision of r in estimating ρ is generally worst when $\rho = 0$
- This margin for error as well as that for three other choices of the unknown true ρ are shown in Figure 8.5.

```
require(Hmisc)
plotCorrPrecision(rho=c(0, .25, .5, .75),
                  n=seq(10, 1000, length=100),
                  ylim=c(0, .4), col=1:4, opts=list(keys='lines'))
abline(h=seq(0, .4, by=0.025),
       v=seq(25, 975, by=25), col=gray(.9))
```

See also stats.stackexchange.com/questions/415131.

8.5.3

Comparing Two r 's

- Rarely appropriate

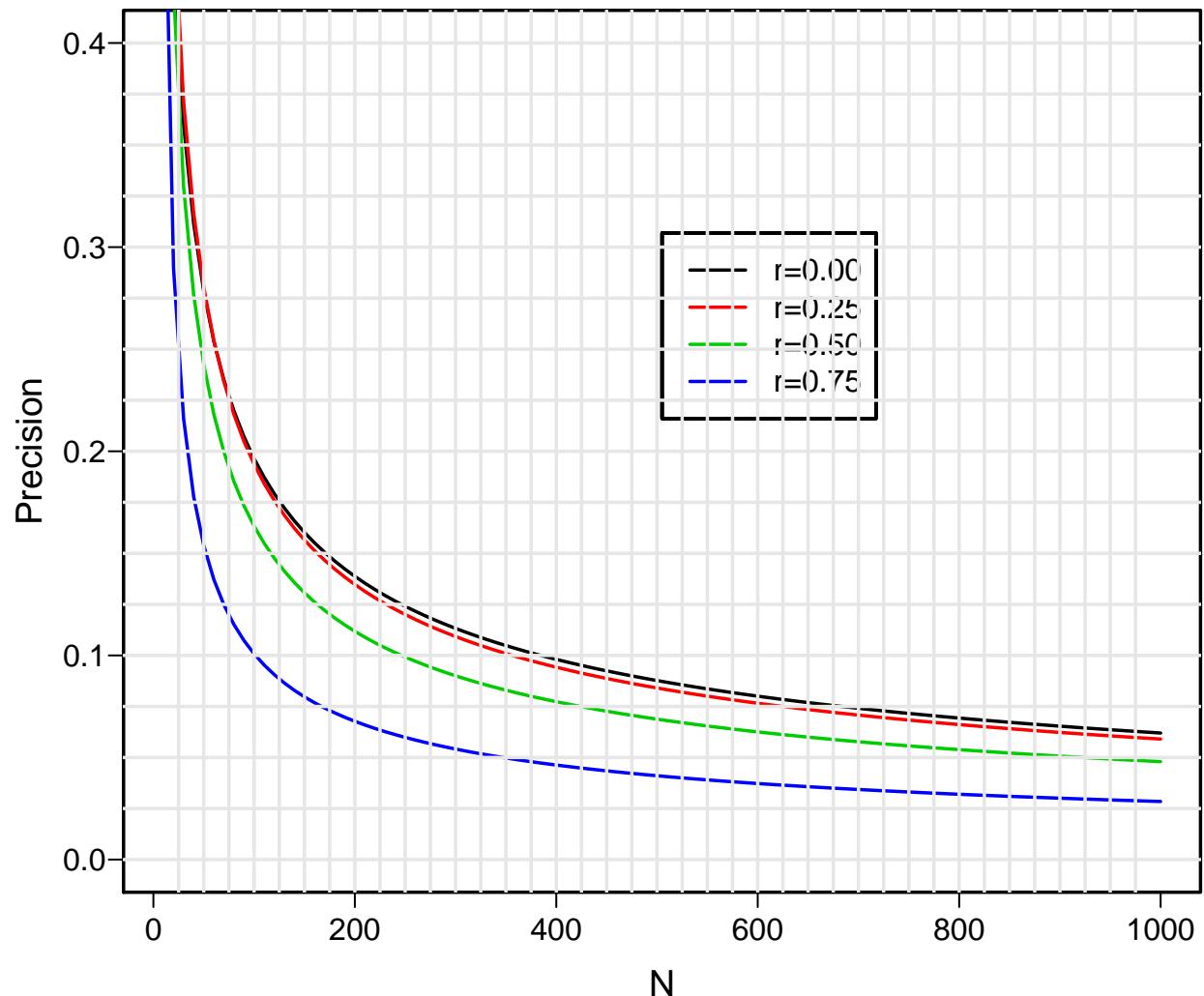


Figure 8.5: Margin for error (length of longer side of asymmetric 0.95 confidence interval) for r in estimating ρ , when $\rho = 0, 0.25, 0.5, 0.75$. Calculations are based on Fisher z transformation of r .

- Two r 's can be the same even though slopes may differ
- Usually better to compare effects on a real scale (slopes)

8.6

Avoiding Overinterpretation

- Often researchers
 - compute many correlations then
 - make a big deal out of the largest observed correlation
- This is *double dipping*: using the same dataset to tell you which features to test, then testing those features
- This is a ranking and selection problem, and the data seldom contain enough information to be reliable in the choices

8.6.1

Simulate Data

- For our analysis experiments, simulate a sample of size 50 from a 10-variate normal distribution with a known correlation matrix
- To specify this correlation matrix take the easy way out: compute an observed correlation matrix from a small sample where all correlations in the population are zero
- The usual sample noise will generate some large observed correlations

```
require(Hmisc)
require(mvtnorm)
```

```
# Get a population correlation matrix by getting sample correlations
# on a random normal sample with N=20 and all true correlations=0
# then pretend these sample correlations were real population values
set.seed(3)
x <- rmvnorm(20, sigma=diag(10))
R <- rcorr(x)$r
# True correlations we will simulate from:
round(R, 2)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	1.00	0.01	0.47	-0.09	-0.31	-0.07	0.14	0.05	0.02	-0.49
[2,]	0.01	1.00	0.48	0.27	0.27	0.14	0.40	-0.17	-0.59	0.60

```
[3,]  0.47  0.48  1.00 -0.11  0.26 -0.31  0.45 -0.12 -0.50 -0.06
[4,] -0.09  0.27 -0.11  1.00  0.42  0.42 -0.07 -0.35  0.16  0.35
[5,] -0.31  0.27  0.26  0.42  1.00 -0.04  0.03 -0.40 -0.25  0.34
[6,] -0.07  0.14 -0.31  0.42 -0.04  1.00  0.19 -0.36  0.08  0.40
[7,]  0.14  0.40  0.45 -0.07  0.03  0.19  1.00 -0.37 -0.65 -0.07
[8,]  0.05 -0.17 -0.12 -0.35 -0.40 -0.36 -0.37  1.00  0.17 -0.24
[9,]  0.02 -0.59 -0.50  0.16 -0.25  0.08 -0.65  0.17  1.00 -0.18
[10,] -0.49  0.60 -0.06  0.35  0.34  0.40 -0.07 -0.24 -0.18  1.00
```

```
# Get a huge sample from a multivariate normal distribution to see
# that it mimics the real correlation matrix R
x <- rmvnorm(50000, sigma=R)
table(round(R - rcorr(x)$r, 2))
```

```
-0.01      0  0.01
 14     76   10
```

```
# Now sample from the population to get our dataset with N=50
x <- rmvnorm(50, sigma=R)
rorig <- rcorr(x)$r
round(rorig, 2)
```

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]  1.00 -0.01  0.50  0.11  0.00 -0.16  0.07 -0.04 -0.04 -0.43
[2,] -0.01  1.00  0.50  0.19  0.43  0.13  0.60 -0.26 -0.76  0.69
[3,]  0.50  0.50  1.00 -0.15  0.45 -0.41  0.52 -0.18 -0.58 -0.01
[4,]  0.11  0.19 -0.15  1.00  0.45  0.30 -0.13 -0.35  0.09  0.14
[5,]  0.00  0.43  0.45  0.45  1.00 -0.12  0.27 -0.53 -0.42  0.20
[6,] -0.16  0.13 -0.41  0.30 -0.12  1.00  0.06 -0.27  0.00  0.51
[7,]  0.07  0.60  0.52 -0.13  0.27  0.06  1.00 -0.57 -0.81  0.19
[8,] -0.04 -0.26 -0.18 -0.35 -0.53 -0.27 -0.57  1.00  0.37 -0.13
[9,] -0.04 -0.76 -0.58  0.09 -0.42  0.00 -0.81  0.37  1.00 -0.41
[10,] -0.43  0.69 -0.01  0.14  0.20  0.51  0.19 -0.13 -0.41  1.00
```

8.6.2

Margin of Error for a Single r

- First compute the margin of error in estimating a single r from $n = 50$
- This is the spacing between r and its lower 0.95 CL or the spacing between r and its upper CL whichever is greatest
- CL based on Fisher's z -transformation described earlier
- Compute this for 4 hypothetical true r : 0 0.25 0.5 0.75

```
r <- (0 : 3) / 4
n <- 50
```

```

zcrit <- qnorm(0.975)
z <- 0.5 * log( (1 + r) / (1 - r))
lo <- z - zcrit/sqrt(n-3)
hi <- z + zcrit/sqrt(n-3)
rlo <- (exp(2*lo)-1)/(exp(2*lo)+1)
rhi <- (exp(2*hi)-1)/(exp(2*hi)+1)
w <- rbind(r=r, 'Margin of Error'=pmax(rhi - r, r - rlo))
prmatrix(round(w, 2), collab=rep(' ', 4))

```

r	0.00	0.25	0.50	0.75
Margin of Error	0.28	0.28	0.24	0.15

- If the true correlation is 0.5, the margin of error in estimating it is ± 0.24 with $n = 50$

8.6.3

Bootstrapping the Correlation Selection Process

- Can use the bootstrap to document the difficulty of the task
- Steps:
 1. form a matrix with N rows and p columns where N is the number of observations and p is the number of variables being correlated with each other
 2. draw a sample of size N from this matrix by sampling its rows
 3. compute all the correlation coefficients as was done previously
 4. re-run the same selection process as was done on the original dataset
 5. repeat 1000 times
 6. examine the distribution of the selections over the 1000 repeats

8.6.4

Bootstrapping Bias in Largest Observed r

- Start with something simpler than ranking all the r s in the matrix: estimate the bias in the largest observed r
- Draw a sample with replacement from rows of the data matrix

- For each of these bootstrap samples find which pair of variables has the highest r
- Track this variable pair and compute r in the original sample
- Compute the dropoff in r from the bootstrap sample to the original sample
- This simulates the *process* used to find the largest r
- Example: use data simulated above with 10 standard normal random variables & known correlation matrix on 50 subjects
- Sample correlation matrix has $\frac{10 \times 9}{2} = 45$ distinct coefficients

```
# Function to retrieve the upper triangle of a symmetric matrix
# ignoring the diagonal terms
up <- function(z) z[upper.tri(z)]
rorigu <- up(rorig)
max(rorigu) # .685 = [2,10] element; 0.604 in population
```

```
[1] 0.6854261
```

```
which.max(rorigu)
```

```
[1] 38
```

```
# is the 38th element in the upper triangle

# Tabulate the difference between sample r estimates and true values
Ru <- up(R)
mean(abs(Ru - rorigu))
```

```
[1] 0.1149512
```

```
table(round(Ru - rorigu, 1))
```

-0.3	-0.2	-0.1	0	0.1	0.2
2	6	7	7	17	6

```
# Repeat the "finding max r" procedure for 1000 bootstrap samples
# Sample from x 1000 times with replacement, each time computing
# a new correlation matrix
samepair <- dropsum <- 0
for(i in 1 : 1000) {
  b <- sample(1 : 50, replace=TRUE)
  xb <- x[b, ] # sample with replacement from rows
  r <- rcorr(xb)$r
  ru <- up(r)
  wmax <- which.max(ru)
  if(wmax == 38) samepair <- samepair + 1
  # Compute correlation for the bootstrap best pair in the original sample
```

```

    origr ← rorigu[wmax]
    # Compute dropoff in best r
    dropoff ← ru[wmax] - origr
    dropsum ← dropsum + dropoff
}
cat('Number of bootstaps selecting the original most correlated pair: ',
    samepair, 'out of 1000', '\n')

```

Number of bootstaps selecting the original most correlated pair: 642 out of 1000

```
cat('Mean dropoff for max r:', round(dropsum / 1000, 3), '\n')
```

Mean dropoff for max r: 0.071

- For our dataset with $n = 50$ we expect that the maximum observed r out of 45 rs is biased high by 0.071
 - Could subtract 0.071 to debias the observed max r although this will worsen its precision

8.6.5

Bootstrapping Ranks of All rs

- Do a more comprehensive assessment that quantifies the difficulty of the task in ranking all the rs
- Quantify uncertainties in the ranks of the original correlation coefficients
- Apparent “winner” was the one receiving the highest ranking q among all rs
- What is the distribution of q over the 1000 bootstrap samples?
- Can easily compute a 0.95 bootstrap nonparametric percentile confidence interval for the true unknown ranking of that feature combination and for ranking all the examined feature correlations
- Bootstrap the correlation matrix and re-rank the coefficients
- For each pair of variables compute the 0.95 confidence interval for the rank of its correlation from among the 45

```

# For each observed correlation compute its rank among 45 distinct pairs
orig.ranks ← rank(up(rorig))
# Sample from x 1000 times with replacement, each time computing
# a new correlation matrix
Rm ← matrix(NA, nrow=1000, ncol=45)
for(i in 1 : 1000) {
  b ← sample(1 : 50, replace=TRUE)
  xb ← x[b, ]
  r ← rcorr(xb)$r
  Rm[i, ] ← rank(up(r))
}
# Over bootstrap correlations compute quantiles of ranks
low ← apply(Rm, 2, quantile, probs=0.025)
high ← apply(Rm, 2, quantile, probs=0.975)
round(cbind('Original Rank'=orig.ranks, Lower=low, Upper=high))

```

	Original	Rank	Lower	Upper
[1,]		22	12	34
[2,]		40	32	45
[3,]		41	30	45
[4,]		28	15	37
[5,]		31	20	38
[6,]		15	8	28
[7,]		24	11	34
[8,]		37	32	43
[9,]		38	32	43
[10,]		39	31	44
[11,]		14	8	27
[12,]		29	15	37
[13,]		9	3	15
[14,]		35	22	43
[15,]		18	10	26
[16,]		26	16	34
[17,]		44	38	45
[18,]		43	34	45
[19,]		16	9	27
[20,]		34	25	38
[21,]		25	14	34
[22,]		19	11	32
[23,]		12	7	21
[24,]		13	7	27
[25,]		10	4	20
[26,]		5	3	10
[27,]		11	5	22
[28,]		4	3	8
[29,]		20	10	33
[30,]		2	1	4
[31,]		3	2	11
[32,]		27	16	36
[33,]		7	4	13
[34,]		23	13	32
[35,]		1	1	2
[36,]		36	28	43
[37,]		6	3	15
[38,]		45	40	45
[39,]		21	12	31

[40,]	30	17	37
[41,]	33	21	39
[42,]	42	34	44
[43,]	32	20	38
[44,]	17	11	28
[45,]	8	3	16

- Highest observed r (rank 45) has 0.95 CI [40, 45]
- Data are consistent with it being the 6th highest
- Smallest observed value (rank 1; most impressive negative correlation) has CI [1, 2] for rank; data consistent with that pair of variables being the best or 2nd
- The r originally ranked 36th has a 0.95 CI of [28, 43] so the data are consistent with it being in the top 3

8.6.6

Monte Carlo Simulation to Get A Rank Interval

- For comparison with the bootstrap, get the frequency distribution of ranks in repeated studies of the apparently highest r in our $n = 50$ study
- Repeated studies will also have $n = 50$ and will be generated from the population correlation matrix
- Recall that original max r was the 38th element of the strung-out r matrix from our sample

```
ranks <- integer(1000)
for(i in 1 : 1000) {
  xs <- rmvnorm(50, sigma=R)
  rsim <- up(rcorr(xs)$r)
  ranks[i] <- rank(rsim)[38]
}
table(ranks) # freqs. of ranks of 38th element in new samples
```

ranks														
34	35	36	37	38	39	40	41	42	43	44	45			
2	1	6	8	10	18	25	38	58	88	152	594			

```
quantile(ranks, c(0.025, 0.975))
```

2.5%	97.5%
38	45

- This interval is a bit wider than the bootstrap interval
- **Note:** both the bootstrap and the Monte Carlo simulated interval would be **far wider** had the number of correlation coefficients estimated been much greater than the sample size
- Example: consider 1000 possible associations with a single Y
- This time the potential predictors X will be independent in the population and they will be conditioned on (held constant over simulations at the original X values)
- The true importance of the X s is 1, 2, ..., 10 for the first 10 and all remaining are irrelevant in the population

```
set.seed(8)
n ← 50
p ← 1000
x ← matrix(rnorm(n * p), ncol=p)
Ey ← x[,1] + 2 * x[,2] + 3 * x[,3] + 4 * x[,4] + 5 * x[,5] + 6 * x[,6] +
    7 * x[,7] + 8 * x[,8] + 9 * x[,9] + 10 * x[,10]
y ← Ey + rnorm(n) * 20
ro ← cor(x, y)
# First 10 correlations and tabulate all others
round(ro[1:10], 2)
```

```
[1] 0.26 -0.16 0.05 0.29 -0.03 0.04 0.28 0.05 0.28 0.57
```

```
table(round(ro[-(1:10)], 1))
```

-0.5	-0.4	-0.3	-0.2	-0.1	0	0.1	0.2	0.3	0.4	0.5
1	6	43	97	234	250	210	120	25	3	1

```
# Find which of the 1000 correlation against  $Y$  is largest
wmax ← which.max(ro)
wmax      # correct according to population
```

```
[1] 10
```

```
ro[wmax]      # original rank = 1000
```

```
[1] 0.5698995
```

```
# Simulate 1000 repeats of sample with new y but keeping x the same
# See how original highest correlating variable ranks among 1000
# correlations in new samples

ranks ← numeric(1000)
for(i in 1 : 1000) {
  ys ← Ey + rnorm(n) * 20
  rs ← cor(x, ys)
  ranks[i] ← rank(rs)[wmax]
}

table(round(ranks, -1)) # round to nearest 10
```

350	400	580	600	610	620	630	660	670	680	690	700	710	720	730	740
1	1	1	1	1	2	1	1	4	2	2	2	1	2	1	
1	760	770	780	790	800	810	820	830	840	850	860	870	880	890	900
15	1	1	2	4	6	2	6	11	3	6	6	9	13	8	15
15	920	930	940	950	960	970	980	990	1000						
15	34	39	41	50	66	126	203	294							

```
quantile(ranks, c(0.025, 0.975))
```

```
2.5%    97.5%
770.85 1000.00
```

```
sum(ranks > 998)
```

```
[1] 139
```

- The apparent winning variable could fairly easily be the 771st largest instead of the 1000th ranked correlation
- The winning variable was in the top two in only 139 out of 1000 simulations
- See Chapter 20 for more ways to quantify limitations of high-dimensional data analysis, including the $p > N$ case

Chapter 9

Introduction to the R `rms` Package: The Linear Model

Some of the purposes of the `rms` package are to

A

- make everyday statistical modeling easier to do
- make modern statistical methods easy to incorporate into everyday work
- make it easy to use the bootstrap to validate models
- provide “model presentation graphics”

9.1

Formula Language and Fitting Function

B

- Statistical formula in S:

```
y ~ x1 + x2 + x3
```

y is modeled as $\alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$.

- *y* is the dependent variable/response/outcome, *x*'s are predictors (independent variables)
- The formula is the first argument to a *fitting* function (just as it is the first argument to a *trellis* graphics function)
- rms (*regression modeling strategies*) package⁴¹ makes many aspects of regression modeling and graphical display of model results easier to do
- rms does a lot of bookkeeping to remember details about the *design matrix* for the model and to use these details in making automatic hypothesis tests, estimates, and plots. The design matrix is the matrix of independent variables after coding them numerically and adding nonlinear and product terms if needed.
- rms package fitting function for ordinary least squares regression (what is often called the *linear model* or *multiple linear regression*): `ols`
- Example:

```
f <- ols(y ~ age + sys.bp, data=mydata)
```

C

- *age* and *sys.bp* are the two predictors (independent variables) assumed to have linear and additive effects (do not interact or have synergism)
- *mydata* is an R *data frame* containing at least three columns for the model's variables
- *f* (the *fit object*) is an R list object, containing coefficients, variances, and many other quantities
- Below, the fit object will be *f* throughout. In practice, use any legal R name, e.g. `fit.full.model`

9.2

Operating on the Fit Object

D

- Regression coefficient estimates may be obtained by any of the methods listed below

```
f$coefficients
f$coef           # abbreviation
coef(f)          # use the coef extractor function
coef(f)[1]        # get intercept
f$coef[2]         # get 2nd coefficient (1st slope)
f$coef['age']     # get coefficient of age
coef(f)['age']    # ditto
```

- But often we use *methods* which do something more interesting with the model fit.

`print(f)` : print coefficients, standard errors, *t*-test, other statistics; can also just type `f` to print

`fitted(f)` : compute \hat{y}

`predict(f, newdata)` : get predicted values, for subjects described in data frame
`newdata`^a

`r <- resid(f)` : compute the vector of n residuals (here, store it in `r`)

`formula(f)` : print the regression formula fitted

`anova(f)` : print ANOVA table for all total and partial effects

`summary(f)` : print estimates partial effects using meaningful changes in predictors

`Predict(f)` : compute predicted values varying a few predictors at a time (convenient for plotting)

`ggplot(p)` : plot partial effects, with predictor ranging over the x -axis, where `p` is the result of `Predict`

`g <- Function(f)` : create an R function that evaluates the analytic form of the fitted function

`nomogram(f)` : draw a nomogram of the model

^aYou can get confidence limits for predicted means or predicted individual responses using the `conf.int` and `conf.type` arguments to `predict`. `predict(f)` without the `newdata` argument yields the same result as `fitted(f)`.

9.3

The rms datadist Function

E

To use Predict, summary, or nomogram in the rms package, you need to let rms first compute summaries of the distributional characteristics of the predictors:

```
dd ← datadist(x1,x2,x3,...)    # generic form
dd ← datadist(age, sys.bp, sex)
dd ← datadist(mydataframe)      # for a whole data frame
options(datadist='dd')          # let rms know where to find
```

Note that the name `dd` can be any name you choose as long as you use the same name in quotes that you specify (unquoted) to the left of `<- datadist(...)`. It is best to invoke `datadist` early in your program before fitting any models. That way the `datadist` information is stored in the `fit` object so the model is self-contained. That allows you to make plots in later sessions without worrying about `datadist`.

`datadist` must be re-run if you add a new predictor or recode an old one. You can update it using for example

```
dd ← datadist(dd, cholesterol, height)
# Adds or replaces cholesterol, height summary stats in dd
```

9.4

Short Example

F
H

Consider the lead exposure dataset from B. Rosner *Fundamentals of Biostatistics* and originally from Landrigan PJ et al, Lancet 1:708-715, March 29, 1975. The study was of psychological and neurologic well-being of children who lived near a lead smelting plant. The primary outcome measures are the Wechsler full-scale IQ score (`iqf`) and the finger-wrist tapping score `maxfwt`. The dataset is available at biostat.mc.vanderbilt.edu/DataSets and can be automatically downloaded and `load()`'d into R using the `Hmisc` package `getHdata` function. For now we just analyze lead exposure levels in 1972 and 1973, age, and `maxfwt`^b.

Note: To easily run all the following commands, open <http://fharrell.com/code/bbr.zip> and then open the file `8-rmsintro.r` contained in the `.zip` file using RStudio. Commands listed in previous sections were not actually executed so they are marked with the R comment symbol (#) and can be ignored.

```
# For an Rmarkdown version of similar analyses see
# https://github.com/harrelfe/rscripts/raw/master/lead-ols.md
require(rms)      # also loads the Hmisc package

getHdata(lead)
# Subset variables just so contents() and describe() output is short
# Override units of measurement to make them legal R expressions
lead ← upData(lead,
              keep=c('ld72', 'ld73', 'age', 'maxfwt'),
              labels=c(age='Age'),
              units=c(age='years', ld72='mg/100*ml', ld73='mg/100*ml'))
```

```
Input object size:      50832 bytes;    39 variables   124 observations
Kept variables  ld72,ld73,age,maxfwt
New object size:     12824 bytes;     4 variables    124 observations
```

```
contents(lead)
```

```
Data frame:lead 124 observations and 4 variables   Maximum # NAs:25
```

	Labels	Units	Storage	NAs
age	Age	years	double	0
ld72	Blood Lead Levels, 1972	mg/100*ml	integer	0
ld73	Blood Lead Levels, 1973	mg/100*ml	integer	0
maxfwt	Maximum mean finger-wrist tapping score		integer	25

```
describe(lead)  #
```

H

^b`maxfwt` might be better analyzed as an ordinal variable but as will be seen by residual plots it is also reasonably considered to be continuous and to satisfy ordinary regression assumptions.

```

lead

 4 Variables      124 Observations
-----
age : Age [years]
  n    missing   distinct     Info      Mean      Gmd      .05      .10
  124        0       73         1     8.935     4.074     3.929     4.333
  .25       .50       .75       .90       .95
  6.167    8.375    12.021    14.000    15.000

lowest : 3.750000 3.833333 3.916667 4.000000 4.166667
highest: 14.250000 15.000000 15.250000 15.416667 15.916667
-----
ld72 : Blood Lead Levels, 1972 [mg/100*ml]
  n    missing   distinct     Info      Mean      Gmd      .05      .10
  124        0       47         0.999    36.16     17.23     18.00     21.00
  .25       .50       .75       .90       .95
  27.00    34.00    43.00     57.00    61.85

lowest : 1 2 10 14 18, highest: 62 64 66 68 99
-----
ld73 : Blood Lead Levels, 1973 [mg/100*ml]
  n    missing   distinct     Info      Mean      Gmd      .05      .10
  124        0       37         0.998    31.71     11.06     18.15     21.00
  .25       .50       .75       .90       .95
  24.00    30.50    37.00     47.00    50.85

lowest : 15 16 18 19 20, highest: 52 53 54 57 58
-----
maxfwt : Maximum mean finger-wrist tapping score
  n    missing   distinct     Info      Mean      Gmd      .05      .10
  99        25       40         0.998    51.96     13.8      33.2      38.0
  .25       .50       .75       .90       .95
  46.0     52.0     59.0      65.0     72.2

lowest : 13 14 23 26 34, highest: 74 76 79 83 84
-----
```

```

dd ← datadist(lead); options(datadist='dd')
dd      # show what datadist computed
|
```

	age	ld72	ld73	maxfwt
Low:effect	6.166667	27.00	24.00	46.0
Adjust to	8.375000	34.00	30.50	52.0
High:effect	12.020833	43.00	37.00	59.0
Low:prediction	3.929167	18.00	18.15	33.2
High:prediction	15.000000	61.85	50.85	72.2
Low	3.750000	1.00	15.00	13.0
High	15.916667	99.00	58.00	84.0

```

# Fit an ordinary linear regression model with 3 predictors assumed linear
f ← ols(maxfwt ~ age + ld72 + ld73, data=lead)
f      # same as print(f)
|
```

```

Frequencies of Missing Values Due to Each Variable
maxfwt    age    ld72    ld73
|
```

```
25      0      0      0
```

Linear Regression Model

```
ols(formula = maxfwt ~ age + ld72 + ld73, data = lead)
```

	Model	Likelihood Ratio Test	Discrimination Indexes		
Obs	99	LR chi2	62.25	R2	0.467
sigma9	5.221	d.f.		R2 adj	0.450
d.f.	95	Pr(> chi2)	0.0000	g	10.104

Residuals

	Min	1Q	Median	3Q	Max
-33.9958	-4.9214	0.7596	5.1106	33.2590	

	Coef	S.E.	t	Pr(> t)
Intercept	34.1059	4.8438	7.04	<0.0001
age	2.6078	0.3231	8.07	<0.0001
ld72	-0.0246	0.0782	-0.31	0.7538
ld73	-0.2390	0.1325	-1.80	0.0744

```
coef(f) # retrieve coefficients
```

	Intercept	age	ld72	ld73
34.1058551	2.6078450	-0.0245978	-0.2389695	

```
specs(f, long=TRUE) # show how parameters are assigned to predictors ,
```

K

```
ols(formula = maxfwt ~ age + ld72 + ld73, data = lead)
```

Units	Label	Assumption	Parameters	d.f.
age years	Age	asis		1
ld72 mg/100*ml	Blood Lead Levels, 1972	asis		1
ld73 mg/100*ml	Blood Lead Levels, 1973	asis		1

	age	ld72	ld73
Low:effect	6.166667	27.00	24.00
Adjust to	8.375000	34.00	30.50
High:effect	12.020833	43.00	37.00
Low:prediction	3.929167	18.00	18.15
High:prediction	15.000000	61.85	50.85
Low	3.750000	1.00	15.00
High	15.916667	99.00	58.00

```
# and predictor distribution summaries driving plots
g <- Function(f) # create an R function that represents the fitted model
# Note that the default values for g's arguments are medians
g
```

L

```
function (age = 8.375, ld72 = 34, ld73 = 30.5)
{
  34.105855 + 2.607845 * age - 0.024597799 * ld72 - 0.23896951 *
```

```
ld73
}
<environment: 0x55e0eb2b2668>

# Estimate mean maxfwt at age 10, .1 quantiles of ld72, ld73 and .9 quantile of
# ld73
# keeping ld72 at .1 quantile
g(age=10, ld72=21, ld73=c(21, 47)) # more exposure in 1973 decreased y by 6
```

```
[1] 54.64939 48.43618
```

```
# Get the same estimates another way but also get std. errors
predict(f, data.frame(age=10, ld72=21, ld73=c(21, 47)), se.fit=TRUE) M
```

```
$linear.predictors
 1         2
54.64939 48.43618

$se.fit
 1         2
1.391858 3.140361
```

9.5

Operating on Residuals

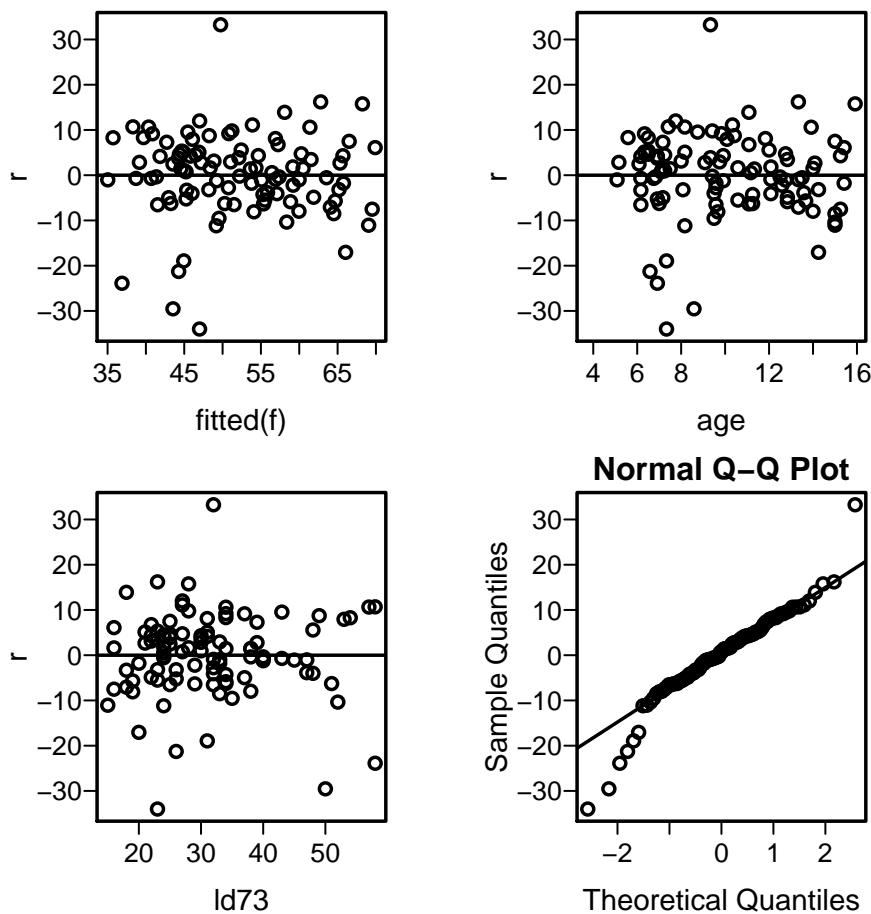


Residuals may be summarized and plotted just like any raw data variable.

N

- To plot residuals vs. each predictor, and to make a q-q plot to check normality of residuals, use these examples:

```
r ← resid(f)
par(mfrow=c(2,2)) # 2x2 matrix of plots
plot(fitted(f), r); abline(h=0) # yhat vs. r
with(lead, plot(age, r)); abline(h=0)
with(lead, plot(ld73, r)); abline(h=0)
qqnorm(r) # linearity indicates normality
qqline(as.numeric(r))
```



9.6

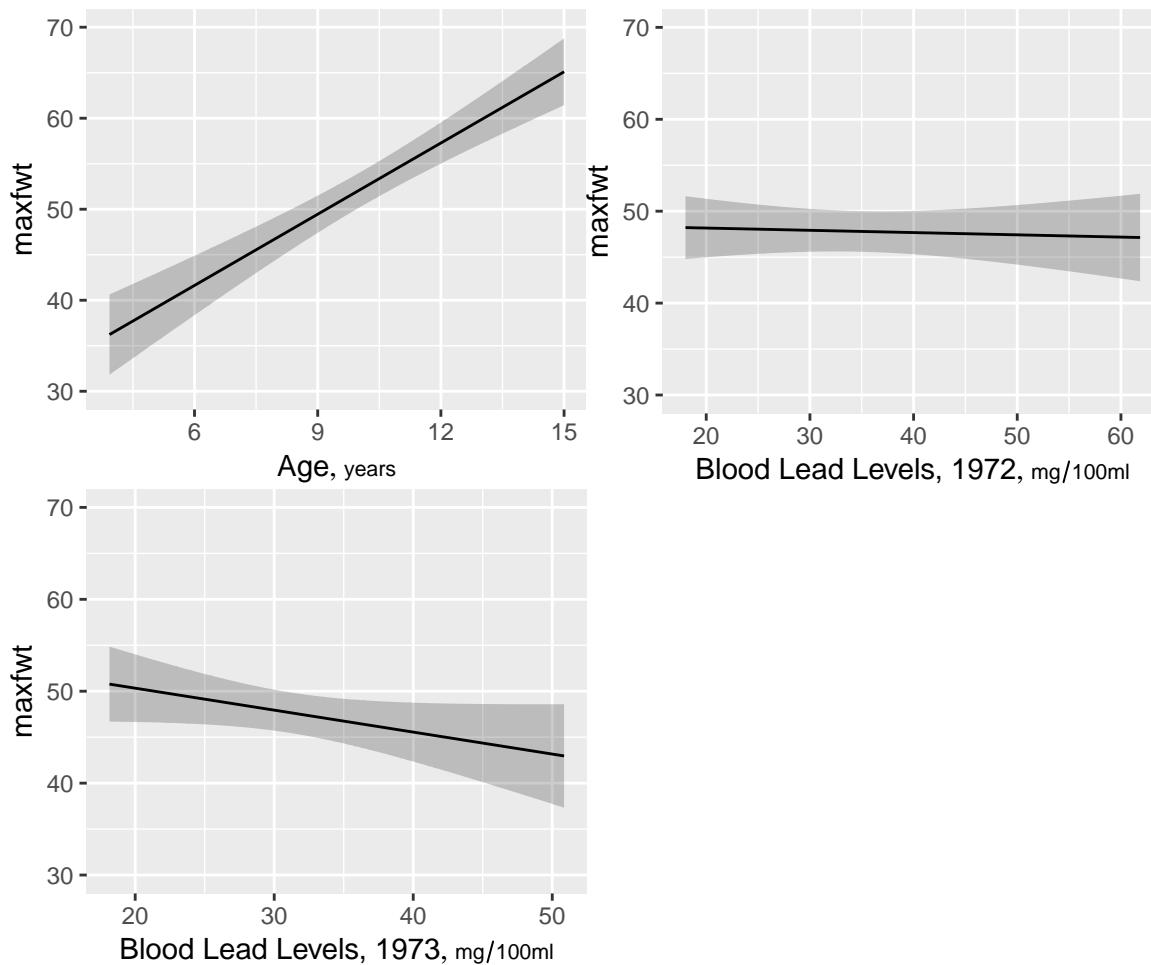
Plotting Partial Effects



- Predict and `ggplot` makes one plot for each predictor
- Predictor is on x -axis, \hat{y} on the y -axis
- Predictors not shown in plot are set to constants
 - median for continuous predictors
 - mode for categorical ones
- For categorical predictor, estimates are shown only at data values
- 0.95 pointwise confidence limits for $\hat{E}(y|x)$ are shown (add `conf.int=FALSE` to `Predict()` to suppress CLs)
- Example:

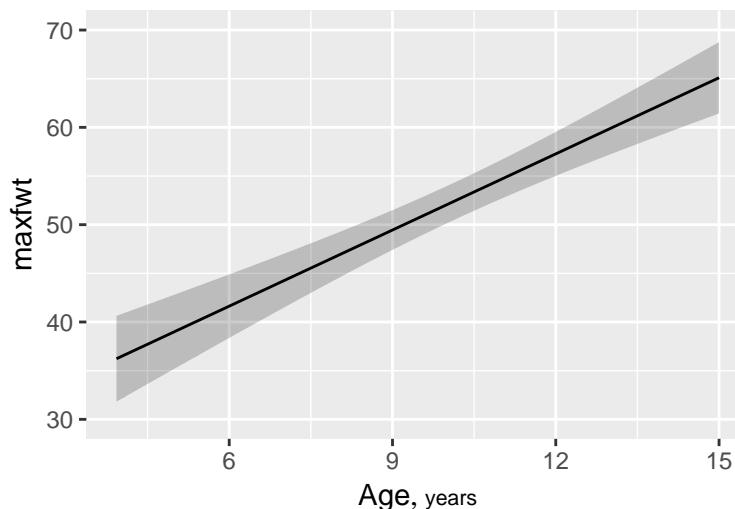
```
ggplot(Predict(f))
```

P



- To take control of which predictors are plotted, or to specify customized options: Q

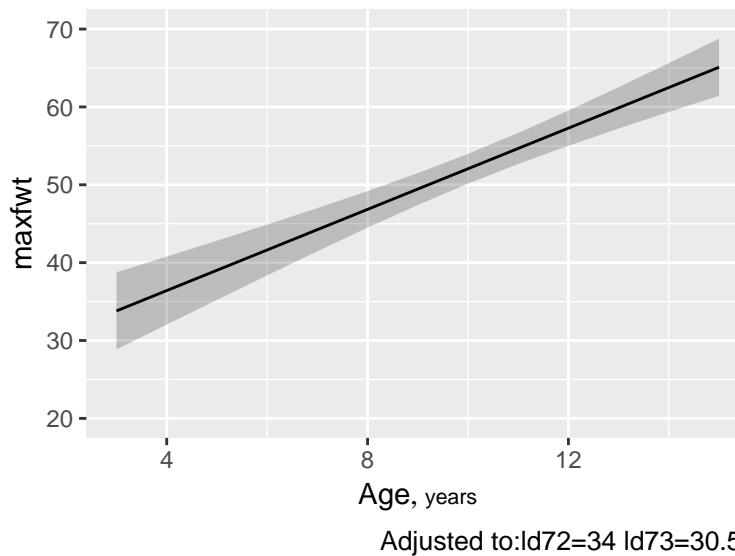
```
ggplot(Predict(f, age))      # plot age effect, using default range,
                             # 10th smallest to 10th largest age
```



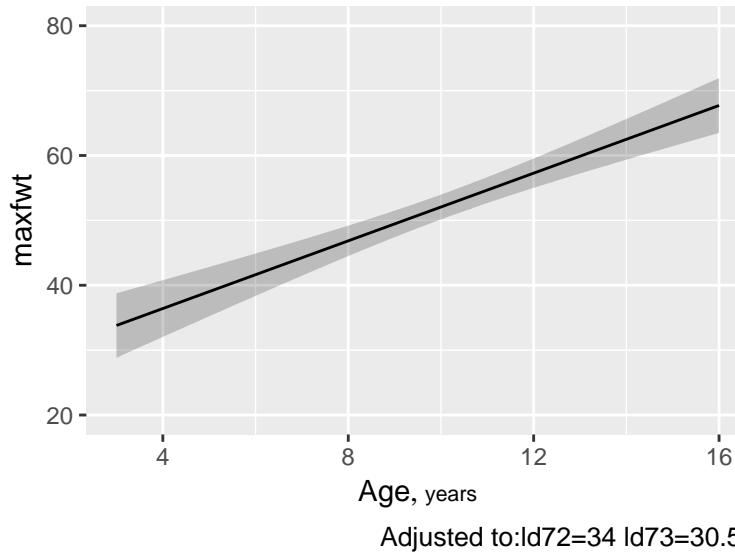
Adjusted to:ld72=34 ld73=30.5

```
ggplot(Predict(f, age=3:15))  # plot age =3, 4, . . . , 15
```

R

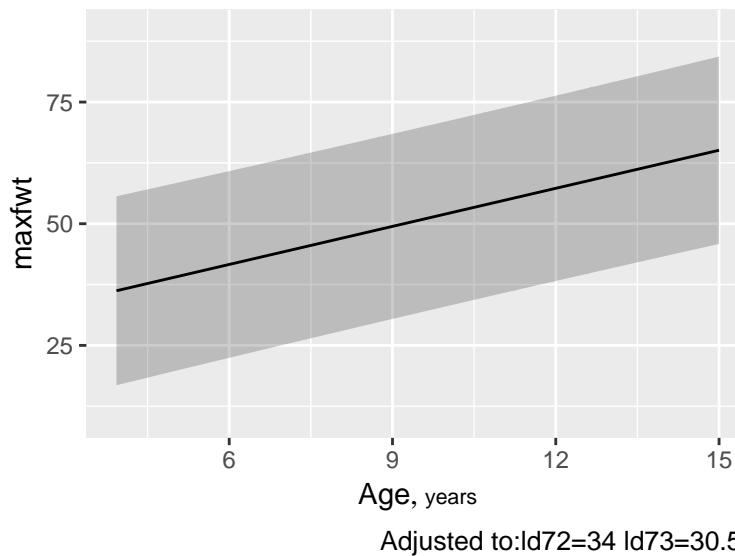


```
ggplot(Predict(f, age=seq(3,16,length=150))) # plot age = 3 - 16, 150 points
```



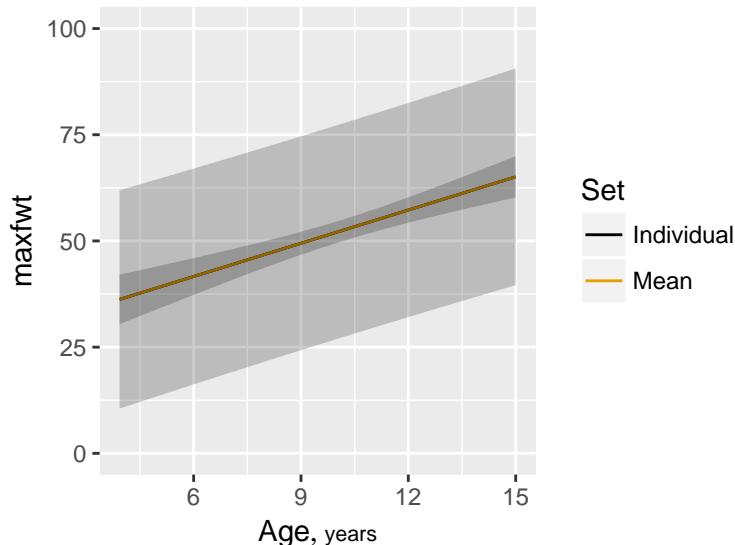
- To get confidence limits for \hat{y} :

```
ggplot(Predict(f, age, conf.type='individual'))
```



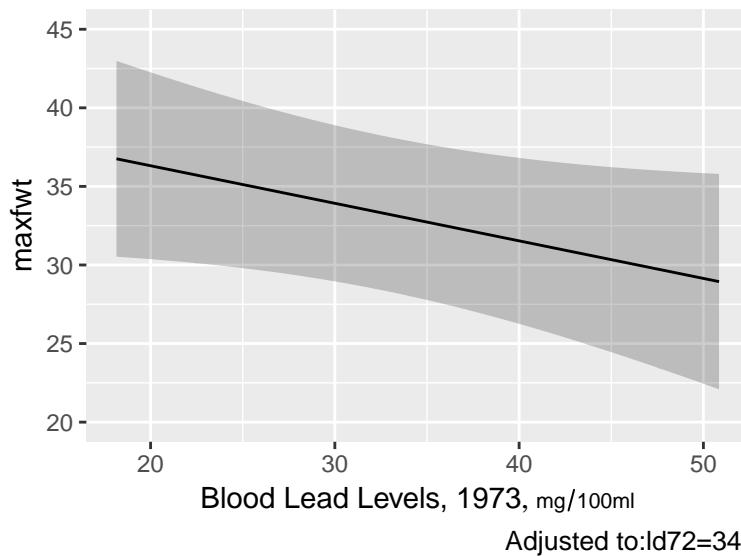
- To show both types of 0.99 confidence limits on one plot:

```
p1 <- Predict(f, age, conf.int=0.99, conf.type='individual')
p2 <- Predict(f, age, conf.int=0.99, conf.type='mean')
p <- rbind(Individual=p1, Mean=p2)
ggplot(p)
```



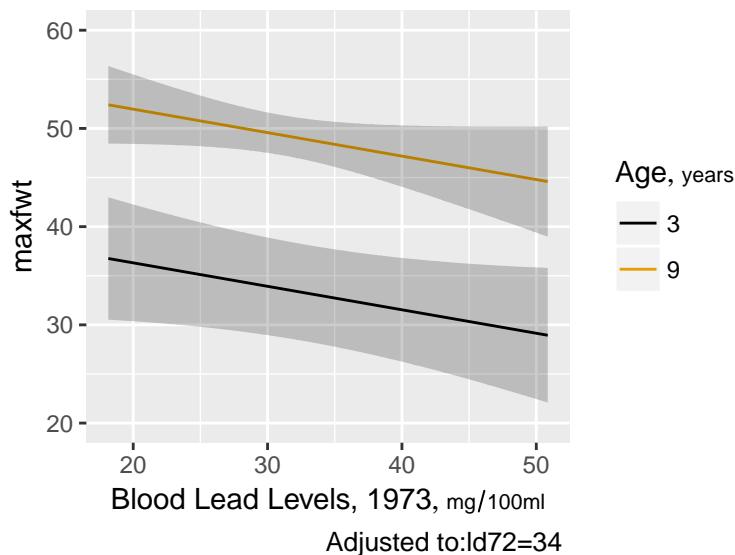
- Non-plotted variables are set to reference values (median and mode by default)
- To control the settings of non-plotted values use e.g.

```
ggplot(Predict(f, ld73, age=3))
```



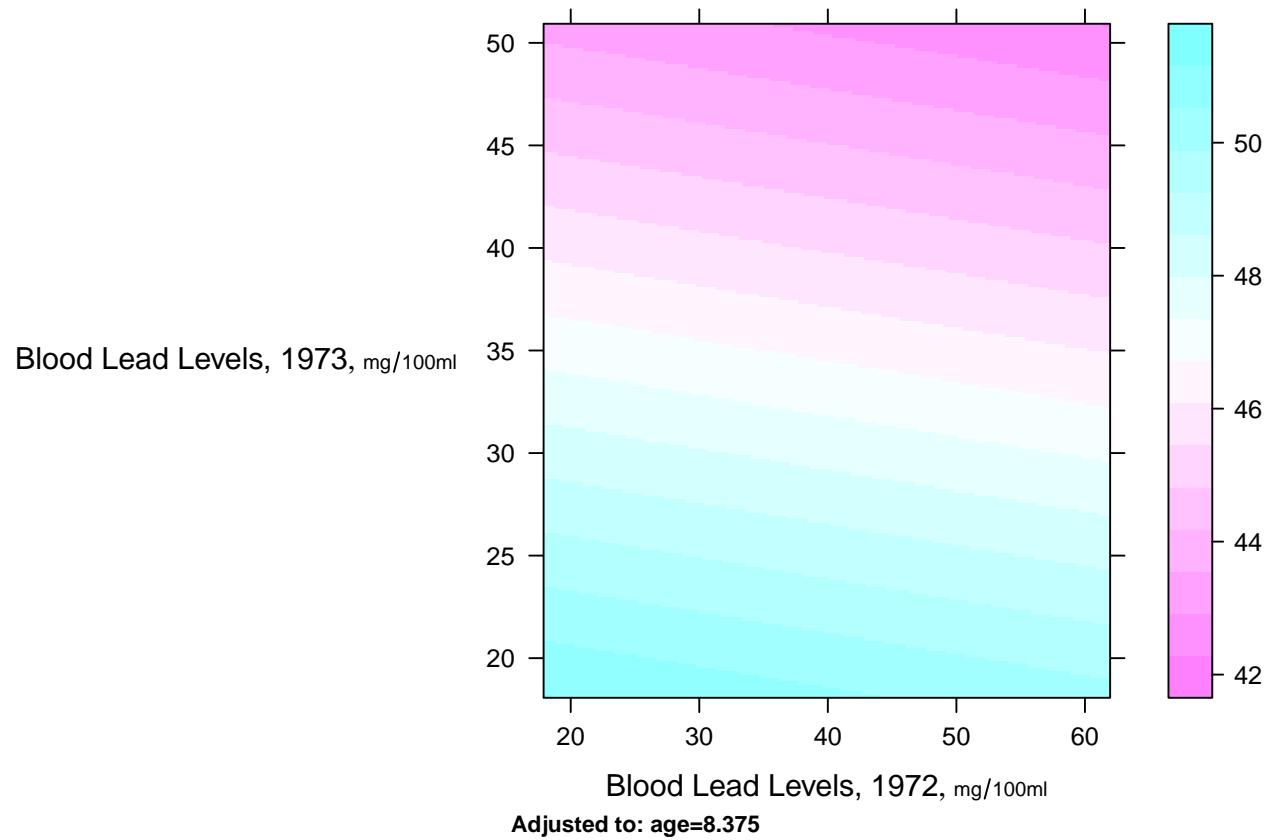
- To make separate lines for two ages:

```
ggplot(Predict(f, ld73, age=c(3,9))) # add ,conf.int=FALSE to suppress
  conf. bands
```



- To plot a 3-d surface for two continuous predictors against \hat{y} ; color coding for predicted mean maxfwt

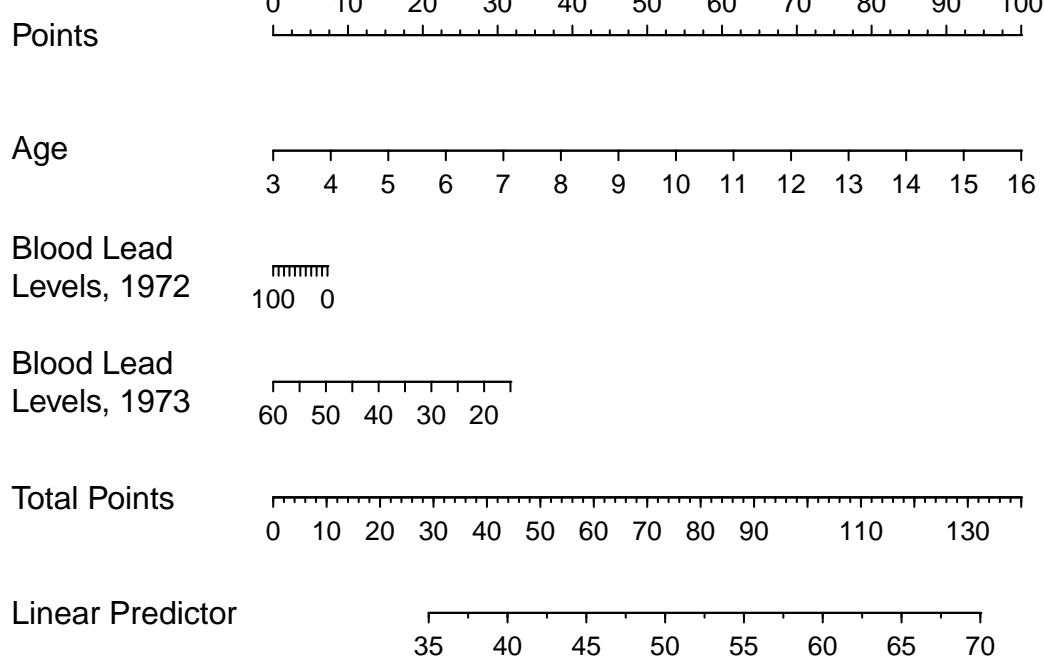
```
bplot(Predict(f, ld72, ld73))
```



9.7

Nomograms: Overall Depiction of Fitted Models

```
plot(nomogram(f))
```



See [this](#) for excellent examples showing how to read such nomograms.

9.7.1

Point Estimates for Partial Effects



The `summary` function can compute point estimates and confidence intervals for effects of individual predictors, holding other predictors to selected constants. The constants you hold other predictors to will only matter when the other predictors interact with the predictor whose effects are being displayed.

How predictors are changed depend on the type of predictor:

A

- Categorical predictors: differences against the reference (most frequent) cell by

default

- Continuous predictors: inter-quartile-range effects by default

The estimated effects depend on the type of model:

B

- ols: differences in means
- logistic models: odds ratios and their logs
- Cox models: hazard ratios and their logs
- quantile regression: differences in quantiles

```
summary(f)          # inter-quartile-range effects
```

	Effects			Response : maxfwt				
Factor	Low	High	Diff.	Effect	S.E.	Lower	0.95	Upper
age	6.1667	12.021	5.8542	15.26700	1.8914	11.5120		19.02200
ld72	27.0000	43.000	16.0000	-0.39356	1.2511	-2.8773		2.09010
ld73	24.0000	37.000	13.0000	-3.10660	1.7222	-6.5255		0.31234


```
summary(f, age=5)  # adjust age to 5 when examining ld72, ld73
```

	Effects			Response : maxfwt				
Factor	Low	High	Diff.	Effect	S.E.	Lower	0.95	Upper
age	6.1667	12.021	5.8542	15.26700	1.8914	11.5120		19.02200
ld72	27.0000	43.000	16.0000	-0.39356	1.2511	-2.8773		2.09010
ld73	24.0000	37.000	13.0000	-3.10660	1.7222	-6.5255		0.31234


```
# (no effect since no interactions in model)
summary(f, ld73=c(20, 40))  # effect of changing ld73 from 20 to 40
```

	Effects			Response : maxfwt				
Factor	Low	High	Diff.	Effect	S.E.	Lower	0.95	Upper
age	6.1667	12.021	5.8542	15.26700	1.8914	11.5120		19.02200
ld72	27.0000	43.000	16.0000	-0.39356	1.2511	-2.8773		2.09010
ld73	20.0000	40.000	20.0000	-4.77940	2.6495	-10.0390		0.48052

When a predictor has a linear effect, its slope is the one-unit change in Y when the predictor increases by one unit. So the following trick can be used to get a confidence interval for a slope: use `summary` to get the confidence interval for the one-unit change:

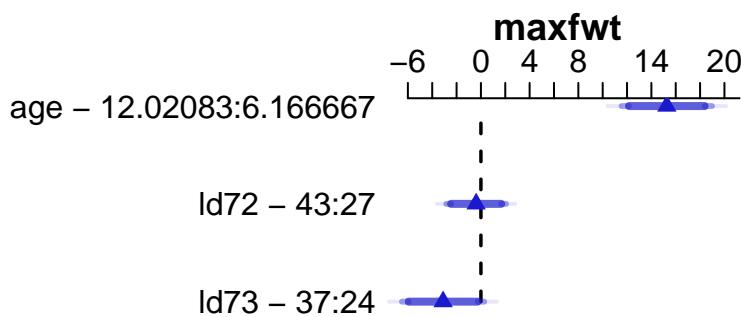
```
summary(f, age=5:6)      # starting age irrelevant since age is linear
```

D

Effects				Response : maxfwt				
Factor	Low	High	Diff.	Effect	S.E.	Lower	0.95	Upper
age	5	6	1	2.60780	0.32308	1.9664	3.24920	
ld72	27	43	16	-0.39356	1.25110	-2.8773	2.09010	
ld73	24	37	13	-3.10660	1.72220	-6.5255	0.31234	

There is a plot method for summary results. By default it shows 0.9, 0.95, and 0.99 confidence limits. E

```
plot(summary(f))
```



9.8

Getting Predicted Values

 F

- Using predict

```
predict(f, data.frame(age=3, ld72=21, ld73=21))
```

```
1  
36.39448
```

```
# must specify all variables in the model
```

```
predict(f, data.frame(age=c(3, 10), ld72=21, ld73=c(21, 47)))
```

```
1          2  
36.39448 48.43618
```

```
# predictions for (3,21,21) and (10,21,47)
```

```
newdat ← expand.grid(age=c(4, 8), ld72=c(21, 47), ld73=c(21, 47))  
newdat
```

	age	ld72	ld73
1	4	21	21
2	8	21	21
3	4	47	21
4	8	47	21
5	4	21	47
6	8	21	47
7	4	47	47
8	8	47	47

```
predict(f, newdat)      # 8 predictions
```

	1	2	3	4	5	6	7	8
39.00232	39.00232	49.43370	38.36278	48.79416	32.78911	43.22049	32.14957	42.58095

```
predict(f, newdat, conf.int=0.95)  # also get CLs for mean
```

G

```
$linear.predictors
 1          2          3          4          5          6          7          8
39.00232 49.43370 38.36278 48.79416 32.78911 43.22049 32.14957 42.58095
```

```
$lower
 1          2          3          4          5          6          7          8
33.97441 46.23595 32.15468 43.94736 25.68920 36.94167 27.17060 38.86475
```

```
$upper
 1          2          3          4          5          6          7          8
44.03023 52.63145 44.57088 53.64095 39.88902 49.49932 37.12854 46.29716
```

```
predict(f, newdat, conf.int=0.95, conf.type='individual')  # CLs for indiv.
```

```
$linear.predictors
 1          2          3          4          5          6          7          8
39.00232 49.43370 38.36278 48.79416 32.78911 43.22049 32.14957 42.58095

$lower
 1          2          3          4          5          6          7          8
19.44127 30.26132 18.46566 29.27888 12.59596 23.30120 12.60105 23.31531

$upper
 1          2          3          4          5          6          7          8
58.56337 68.60609 58.25989 68.30944 52.98227 63.13979 51.69810 61.84659
```

See also gendata.

- The brute-force way

```
# Model is b1 + b2*age + b3*ld72 + b4*ld73
b <- coef(f)
# For 3 year old with both lead exposures 21
b[1] + b[2]*3 + b[3]*21 + b[4]*21
```

Intercept
36.39448

- Using Function function

H

```
g <- Function(f)
g(age=c(3, 8), ld72=21, ld73=21)      # 2 predictions
```

[1] 36.39448 49.43370

```
g(age=3)                      # 3 year old at median ld72, ld73
```

[1] 33.80449

9.9

ANOVA



- Use `anova(fitobject)` to get all total effects and individual partial effects
- Use `anova(f, age, sex)` to get combined partial effects of `age` and `sex`, for example
- Store result of `anova` in an object in you want to print it various ways, or to plot it:

```
an <- anova(f)
an                                # same as print(an)
```

Analysis of Variance						Response: maxfwt
Factor	d.f.	Partial SS	MS	F	P	
age	1	5907.535742	5907.535742	65.15	<.0001	
ld72	1	8.972994	8.972994	0.10	0.7538	
ld73	1	295.044370	295.044370	3.25	0.0744	
REGRESSION	3	7540.087710	2513.362570	27.72	<.0001	
ERROR	95	8613.750674	90.671060			

```
print(an, 'names')      # print names of variables being tested
```

Analysis of Variance						Response: maxfwt
Factor	d.f.	Partial SS	MS	F	P	Tested
age	1	5907.535742	5907.535742	65.15	<.0001	age
ld72	1	8.972994	8.972994	0.10	0.7538	ld72
ld73	1	295.044370	295.044370	3.25	0.0744	ld73
REGRESSION	3	7540.087710	2513.362570	27.72	<.0001	age, ld72, ld73
ERROR	95	8613.750674	90.671060			

```
print(an, 'subscripts') # print subscripts in coef(f) (ignoring
```

J

Analysis of Variance						Response: maxfwt
Factor	d.f.	Partial SS	MS	F	P	Tested
age	1	5907.535742	5907.535742	65.15	<.0001	1
ld72	1	8.972994	8.972994	0.10	0.7538	2
ld73	1	295.044370	295.044370	3.25	0.0744	3
REGRESSION	3	7540.087710	2513.362570	27.72	<.0001	1-3
ERROR	95	8613.750674	90.671060			

Subscripts correspond to:
[1] age ld72 ld73

```
# the intercept) being tested
print(an, 'dots')      # a dot in each position being tested
```

```
Analysis of Variance           Response: maxfwt
Factor      d.f. Partial SS   MS      F      P     Tested
age          1    5907.535742 5907.535742 65.15 <.0001 .
ld72         1     8.972994   8.972994   0.10  0.7538 .
ld73         1    295.044370  295.044370   3.25  0.0744 .
REGRESSION   3    7540.087710 2513.362570 27.72 <.0001 ...
ERROR        95    8613.750674   90.671060

Subscripts correspond to:
[1] age  ld72  ld73
```

```
anova(f, ld72, ld73)  # combine effects into a 2 d.f. test
```

K

```
Analysis of Variance           Response: maxfwt
Factor      d.f. Partial SS   MS      F      P
ld72         1     8.972994   8.972994   0.10  0.7538
ld73         1    295.044370  295.044370   3.25  0.0744
REGRESSION   2    747.283558 373.641779  4.12  0.0192
ERROR        95    8613.750674   90.671060
```

Chapter 10

Simple and Multiple Regression Models

Background



Regression models are used for

- hypothesis testing
- estimation
- prediction
- increasing power and precision for assessing the effect of one variable by adjusting for other variables that partially explain the outcome variable Y
- confounder adjustment—getting adjusted estimates of effects
- checking that existing summary scores (e.g., BMI) adequately summarize their component variables
 - fit a model with log height and log weight and see if ratio of coefficients is -2
- determining whether change, average, or most recent measurement should be emphasized
 - fit a model containing body weight measured 1y ago and at time of treatment initiation
 - if simple change score is an adequate summary of the two weights, the ratio of their coefficients will be about -1

- if most recent weight is all-important, coefficient for weight 1y ago will be very small
- developing new summary scores guided by predicting outcomes

Example

- Observational study of patients receiving treatments A and B
- Females are more likely to receive treatment B → need to adjust for sex
- Regression approach: fit a model with covariates (predictors) treatment and sex; treatment effect is adjusted for sex
- Stratification approach: for males estimate the B-A difference and do likewise for females
Average of the two differences is adjusted for sex

Now instead of sex being the relevant adjustment variable suppose it is age, and older patients tend to get treatment B

- Regression approach: fit a model with treatment and age
Treatment effect attempts to estimate the B-A difference at any chosen (fixed; conditioned on) age
- Stratification approach:
 - divide age into quintiles
 - within each quintile compute the B-A difference
 - average these differences to get an almost age-adjusted treatment effect
 - problem with residual heterogeneity of age within quintiles, especially at outer quintiles which are wider
- Matching approach:
 - for each patient on A find a patient on B within 2y of same age

- if no match exists discard the A patient
- don't use the same B patient again
- discard B patients who were not needed to match an A
- do a matched pairs analysis, e.g. paired t -test
- sample size is reduced $\rightarrow \downarrow$ power

10.1

Stratification vs. Matching vs. Regression



reg-alt

- Some ways to hold one variable x_1 constant when estimating the effect of another variable x_2 (*covariate adjustment*):
 - experimental manipulation of x_1
 - stratify the data on x_1 and for each stratum analyze the relationship between x_2 and Y
 - form matched sets of observations on the basis of x_1 and use a statistical method applicable to matched data
 - use a regression model to estimate the joint effects x_1 and x_2 have on Y ; the estimate of the x_2 effect in the context of this model is essentially the x_2 relationship on Y' where Y' is Y after the x_1 effect is subtracted from it
- Stratification and matching are not effective when x_1 is continuous or there are many x 's to hold constant
- Matching may be useful *before* data acquisition is complete or when sample is too small to allow for regression adjustment for > 1 variable
- Matching after the study is completed usually results in discarding a large number of observations that would have been excellent matches
- Methods that discard information lose power and precision, and the observations discarded are arbitrary, damaging the study's reproducibility
- Most matching methods depend on the row order of observations, again putting reproducibility into question
- There is no principled unique statistical approach to analysis of matched data
- All this points to many advantages of regression adjustment

Stratification and matching can be used to adjust for a small set of variables when

assessing the association between a target variable and the outcome. Neither stratification nor matching are satisfactory when there are many adjustment variables or any of them are continuous. Crude stratification is often used in randomized trials to ensure that randomization stays balanced within subsets of subjects (e.g., males/females, clinical sites). Matching is an effective way to save resources **before** a study is done. For example, with a rare outcome one might sample all the cases and only twice as many controls as cases, matching controls to cases on age within a small tolerance such as 2 years. But once data are collected, matching is arbitrary, ineffective, and wasteful, and there are no principled unique statistical methods for analyzing matched data. For example if one wants to adjust for age via matching, consider these data:

Group	Ages
Exposed	30 35 40 42
Unexposed	29 42 41 42

The matched sets may be constructed as follows:

Set	Data
1	E30 U29
2	E40 U41
3	E42 U42a

U42a refers to the first 42 year old in the unexposed group. There is no match for E35. U42b was not used even though she was perfect match for E42.

1. Matching failed to interpolate for age 35; entire analysis must be declared as conditional on age not in the interval [36, 39]
2. $n \downarrow$ by discarding observations that are easy to match (when the observations they are easily matched to were already matched)
3. Majority of matching algorithms are dependent on the row order of the dataset being analyzed so reproducibility is in question

These problems combine to make post-sample-collection matching unsatisfactory from a scientific standpoint^a.

[Gelman](#) has a nice chapter on matching.

^aAny method that discards already available information should be thought of as unscientific.

10.2

Purposes of Statistical Models

Statisticians, like artists, have the bad habit of falling in love with their models.

George E. P. Box

Most folk behave and thoughtlessly believe that the objective of analysis with statistical tools is to find / identify features in the data—period.

They assume and have too much faith that (1) the data effectively reflect nature and not noise, and (2) that statistical tools can ‘auto-magically’ divine causal relations.

They do not acknowledge that the objective of analysis should be to find interesting features of nature (which to the extent that they indicate causal effects will be reproducible), represent them well, and use these features to make reliable decisions about future cases/situations.

Drew Levy
Genentech
March 2015

Too often the *p*-value ‘satisfices’, for the intent of the latter.

I also think that hypothesis testing and effect estimation are really forms of ersatz prediction—ways of identifying and singling out a factor that allows them to make prediction in a simplistic manner (again satisfying for cognitive ease). Therefore full, well formulated prediction modeling is to be preferred to achieve this unacknowledged goal on these grounds.

- Hypothesis testing
 - Test for no association (correlation) of a predictor (independent variable) and a response or dependent variable (unadjusted test) or test for no association of predictor and response after adjusting for the effects of other predictors
- Estimation
 - Estimate the shape and magnitude of the relationship between a single predictor (independent) variable and a response (dependent) variable
 - Estimate the effect on the response variable of changing a predictor from one value to another
- Prediction
 - Predicting response tendencies, e.g., long-term average response as a function of predictors
 - Predicting responses of individual subjects

10.3

Advantages of Modeling

Even when only testing H_0 a model based approach has advantages:

reg-advantage

- Permutation and rank tests not as useful for estimation
- Cannot readily be extended to cluster sampling or repeated measurements
- Models generalize tests
 - 2-sample t -test, ANOVA → multiple linear regression
 - Wilcoxon, Kruskal-Wallis, Spearman → proportional odds ordinal logistic model
 - log-rank → Cox
- Models not only allow for multiplicity adjustment but for shrinkage of estimates
 - Statisticians comfortable with P -value adjustment but fail to recognize that the difference between the most different treatments is badly biased

Statistical estimation is usually model-based

- Relative effect of increasing cholesterol from 200 to 250 mg/dl on hazard of death, holding other risk factors constant
- Adjustment depends on how other risk factors relate to outcome
- Usually interested in adjusted (partial) effects, not unadjusted (marginal or crude) effects

10.4

Nonparametric Regression

- Estimate tendency (mean or median) of Y as a function of X
- Few assumptions
- Especially handy when there is a single X
- Plotted trend line may be the final result of the analysis
- Simplest smoother: moving average

$X:$	1	2	3	5	8
$Y:$	2.1	3.8	5.7	11.1	17.2

$$\hat{E}(Y|X=2) = \frac{2.1 + 3.8 + 5.7}{3}$$

$$\hat{E}(Y|X=\frac{2+3+5}{3}) = \frac{3.8 + 5.7 + 11.1}{3}$$

- overlap OK
- problem in estimating $E(Y)$ at outer X -values
- estimates very sensitive to bin width
- Moving linear regression far superior to moving avg. (moving flat line)
- Cleveland's moving linear regression smoother *loess* (locally weighted least squares) is the most popular smoother
- Example: differential diagnosis of acute bacterial meningitis vs. viral meningitis

```
require(Hmisc)
getHdata(abm)      # Loads data frame ABM (note case)
with(ABM, {
  glratio <- gl / bloodgl
  tpolys <- polys * whites / 100
  plsmo(tpolys, glratio, xlab='Total Polymorphs in CSF',
         ylab='CSF/Blood Glucose Ratio',      # Fig. 10.1
```

```

xlim=quantile(tpolys, c(.05,.95), na.rm=TRUE),
ylim=quantile(glratio, c(.05,.95), na.rm=TRUE))
scat1d(tpolys); scat1d(glratio, side=4) }

```

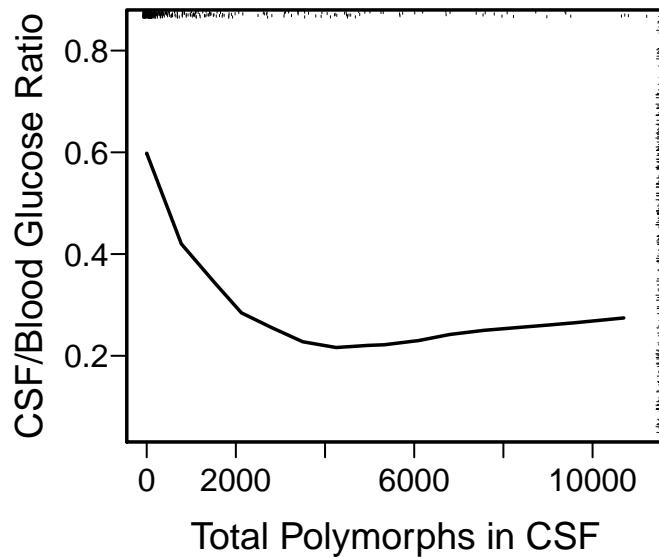


Figure 10.1: loess nonparametric smoother relating CSF:blood glucose ratio to total CSF polymorph count in patients with either bacterial or viral meningitis. Rug plot on axes plots indicate raw data values.

```

with(ABM, {
  plsмо(age, abm, 'supsmu', bass=7,      # Fig. 10.2
        xlab='Age at Admission, Years',
        ylab='Proportion Bacterial Meningitis')
  scat1d(age) })

```

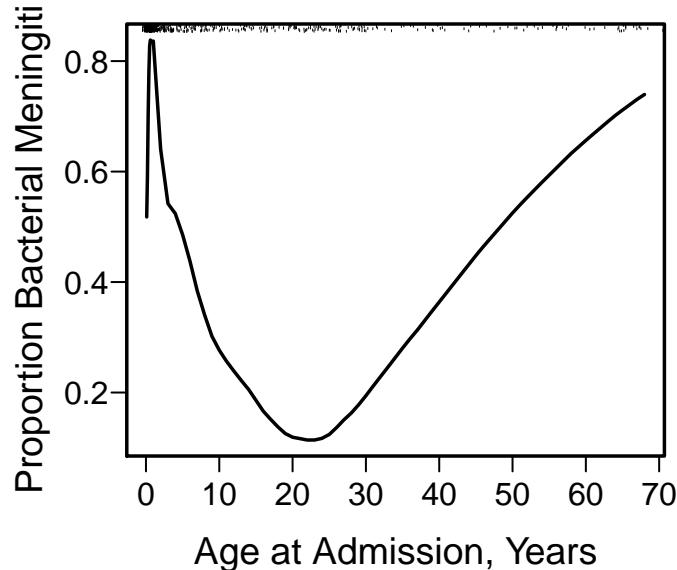


Figure 10.2: "Super smoother" relating age to the probability of bacterial meningitis given a patient has bacterial or viral meningitis, with a rug plot showing the age distribution.

10.5

Simple Linear Regression

10.5.1

Notation

ABD17-17.7,17.10-1



reg-simple

A

- y : random variable representing response variable
- x : random variable representing independent variable (subject descriptor, predictor, covariate)
 - conditioned upon
 - treating as constants, measured without error
- What does conditioning mean?
 - holding constant
 - subsetting on
 - slicing scatterplot vertically

```
n <- 100
set.seed(13)
x <- round(rnorm(n, .5, .25), 1)
y <- x + rnorm(n, 0, .1)
r <- c(-.2, 1.2)
plot(x, y, axes=FALSE, xlim=r, ylim=r, xlab=expression(x), ylab=expression(y))
axis(1, at=r, labels=FALSE)      # Fig. 10.3
axis(2, at=r, labels=FALSE)
abline(a=0,b=1)
histSpike(y, side=2, add=TRUE)
abline(v=.6, lty=2)
```

- $E(y|x)$: population expected value or long-run average of y conditioned on the value of x
 Example: population average blood pressure for a 30-year old
- α : y -intercept

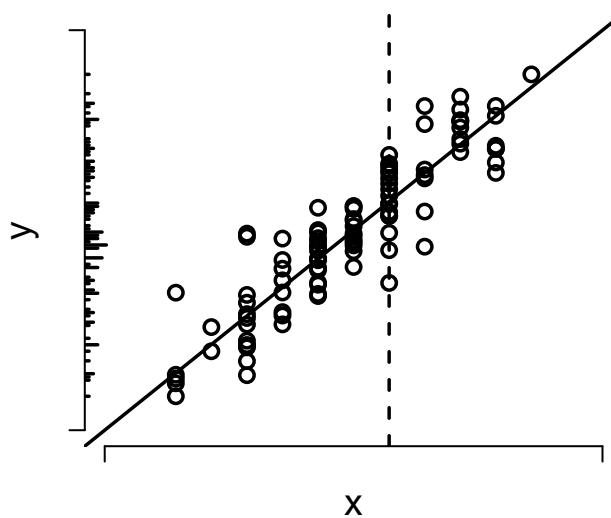


Figure 10.3: Data from a sample of $n = 100$ points along with population linear regression line. The x variable is discrete. The conditional distribution of $y|x$ can be thought of as a vertical slice at x . The unconditional distribution of y is shown on the y -axis. To envision the conditional normal distributions assumed for the underlying population, think of a bell-shaped curve coming out of the page, with its base along one of the vertical lines of points. The equal variance assumption dictates that the series of Gaussian curves for all the different x s have equal variances.

- β : slope of y on x ($\frac{\Delta y}{\Delta x}$)

Simple linear regression is used when

B

- Only two variables are of interest
- One variable is a response and one a predictor
- No adjustment is needed for confounding or other between-subject variation
- The investigator is interested in assessing the strength of the relationship between x and y in real data units, or in predicting y from x
- A linear relationship is assumed (why assume this? why not use nonparametric regression?)
- Not when one only needs to test for association (use Spearman's ρ rank correlation) or estimate a correlation index

10.5.2

Two Ways of Stating the Model

 C

- $E(y|x) = \alpha + \beta x$
- $y = \alpha + \beta x + e$
 e is a random error (residual) representing variation between subjects in y even if x is constant, e.g. variation in blood pressure for patients of the same age

10.5.3

Assumptions, If Inference Needed

D

- Conditional on x , y is normal with mean $\alpha + \beta x$ and constant variance σ^2 , **or:**
- e is normal with mean 0 and constant variance σ^2
- $E(y|x) = E(\alpha + \beta x + e) = \alpha + \beta x + E(e)$,
 $E(e) = 0$.
- Observations are independent

10.5.4

How Can α and β be Estimated?

E

- Need a criterion for what are good estimates
- **One** criterion is to choose values of the two parameters that minimize the sum of squared errors in predicting individual subject responses
- Let a, b be guesses for α, β
- Sample of size n : $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- $SSE = \sum_{i=1}^n (y_i - a - bx_i)^2$

- Values that minimize SSE are *least squares estimates*
- These are obtained from

F

$$L_{xx} = \sum(x_i - \bar{x})^2 \quad L_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta} = b = \frac{L_{xy}}{L_{xx}} \quad \hat{\alpha} = a = \bar{y} - b\bar{x}$$

- Note: A term from L_{xy} will be positive when x and y are concordant in terms of both being above their means or both being below their means.
- Least squares estimates are optimal if
 1. the residuals truly come from a normal distribution
 2. the residuals all have the same variance
 3. the model is correctly specified, i.e., linearity holds

- Demonstration:



```
require(Hmisc)
getRs('demoLeastSquares.r')    # will load code into RStudio script editor
                                # click the Source button to run and follow
                                # instructions in console window
getRs('demoLeastSquares.r', put='source')  # if not using RStudio
```

10.5.5

Inference about Parameters



G

- Estimated residual: $d = y - \hat{y}$
- d large if line was not the proper fit to the data or if there is large variability across subjects for the same x
- Beware of that many authors combine both components when using the terms *goodness of fit* and *lack of fit*
- Might be better to think of lack of fit as being due to a structural defect in the model (e.g., nonlinearity)

- $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

$$SSR = SST - SSE$$

- SS increases in proportion to n

- Mean squares: normalized for d.f.: $\frac{SS}{d.f.(SS)}$

- $MSR = SSR/p$, p = no. of parameters besides intercept (here, 1)

$$MSE = SSE/(n - p - 1) \text{ (sample conditional variance of } y\text{)}$$

$$MST = SST/(n - 1) \text{ (sample unconditional variance of } y\text{)}$$

- Brief review of ordinary ANOVA (analysis of variance):

- Generalizes 2-sample t -test to > 2 groups

- SSR is SS between treatment means

- SSE is SS within treatments, summed over treatments

- ANOVA Table for Regression

H

Source	d.f.	SS	MS	F
Regression	p	SSR	$MSR = SSR/p$	MSR/MSE
Error	$n - p - 1$	SSE	$MSE = SSE/(n - p - 1)$	
Total	$n - 1$	SST	$MST = SST/(n - 1)$	

- Statistical evidence for large values of β can be summarized by $F = \frac{MSR}{MSE}$

- Has F distribution with p and $n - p - 1$ d.f.

- Large values $\rightarrow |\beta|$ large

10.5.6

Estimating σ , S.E. of $\hat{\beta}$; t-test

- $s_{y|x}^2 = \hat{\sigma}^2 = MSE = \widehat{Var}(y|x) = \widehat{Var}(e)$

- $\widehat{se}(b) = s_{y \cdot x} / L_{xx}^{\frac{1}{2}}$
- $t = b / \widehat{se}(b), n - p - 1$ d.f.
- $t^2 \equiv F$ when $p = 1$
- $t_{n-2} \equiv \sqrt{F_{1,n-2}}$
- t identical to 2-sample t -test (x has two values)
- If x takes on only the values 0 and 1, b equals \bar{y} when $x = 1$ minus \bar{y} when $x = 0$

10.5.7

Interval Estimation



- 2-sided $1 - \alpha$ CI for β : $b \pm t_{n-2,1-\alpha/2} \widehat{se}(b)$
- CI for *predictions* depend on what you want to predict even though \hat{y} estimates both y ^b and $E(y|x)$
- Notation for these two goals: \hat{y} and $\hat{E}(y|x)$
 - Predicting y with \hat{y} : K

$$\widehat{s.e.}(\hat{y}) = s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}}}$$

Note: This s.e. $\rightarrow s_{y \cdot x}$ as $n \rightarrow \infty$

 - * As $n \rightarrow \infty$, $\widehat{s.e.}(\hat{y}) \rightarrow s_{y \cdot x}$ which is > 0
 - * Fits with thinking of a predicted individual value being the predicted mean plus a randomly chosen residual (the former has SD going to zero as $n \rightarrow \infty$ and the latter has SD of $s_{y \cdot x}$)
 - * This is a valid predicted individual value but not the lowest mean squared error prediction which would be just to predict at the middle (the mean)
 - Predicting $\hat{E}(y|x)$ with \hat{y} : L

^bWith a normal distribution, the least dangerous guess for an individual y is the estimated mean of y .

$$\widehat{s.e.}(\hat{E}(y|x)) = s_{y,x} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{L_{xx}}} \text{ See footnote}^c$$

Note: This s.e. shrinks to 0 as $n \rightarrow \infty$

- $1 - \alpha$ 2-sided CI for either one:

$$\hat{y} \pm t_{n-p-1, 1-\alpha/2} \widehat{s.e.}$$

- Wide CI (large $\widehat{s.e.}$) due to:

- small n

- large σ^2

- being far from the data center (\bar{x})

- Example usages:

- Is a child of age x smaller than predicted for her age?

- Use $s.e.(\hat{y})$

- What is the best estimate of the population mean blood pressure for patients on treatment A ?

- Use $s.e.(\hat{E}(y|x))$

- Example pointwise 0.95 confidence bands:

x	1	3	5	6	7	9	11
$y:$	5	10	70	58	85	89	135

```
require(rms)
x1 <- c(1, 3, 5, 6, 7, 9, 11)
y <- c(5, 10, 70, 58, 85, 89, 135)
dd <- datadist(x1, n.unique=5); options(datadist='dd')
f <- ols(y ~ x1)
p1 <- Predict(f, x1=seq(1,11,length=100), conf.type='mean')
p2 <- Predict(f, x1=seq(1,11,length=100), conf.type='individual')
p <- rbind(Mean=p1, Individual=p2)
ggplot(p, legend.position='none') +      # Fig. 10.4
      geom_point(aes(x1, y), data=data.frame(x1, y, .set.=''))
```

M

^c n here is the grand total number of observations because we are borrowing information about neighboring x -points, i.e., using interpolation. If we didn't assume anything and just computed mean y at each separate x , the standard error would instead be estimated by $s_{y,x} \sqrt{\frac{1}{m}}$, where m is the number of original observations with x exactly equal to the x for which we are obtaining the prediction. The latter s.e. is much larger than the one from the linear model.

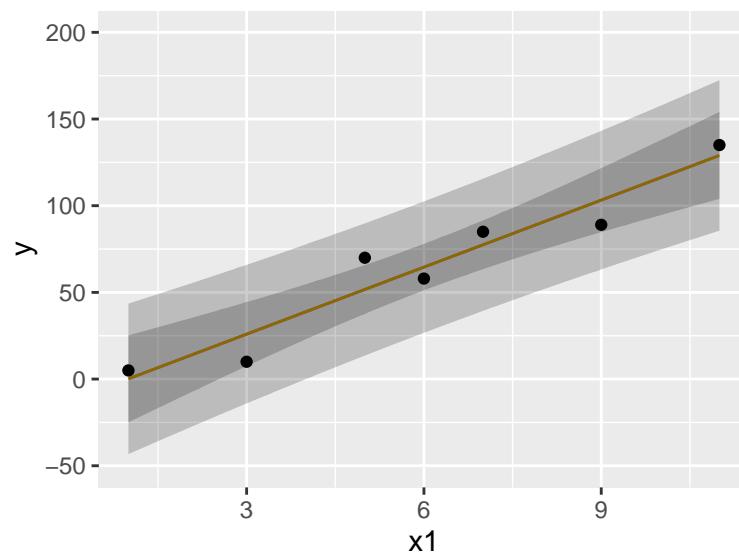


Figure 10.4: Pointwise 0.95 confidence intervals for \hat{y} (wider bands) and $\hat{E}(y|x)$ (narrower bands).

10.5.8

Assessing Goodness of Fit



reg-simple-gof
N

Assumptions:

1. Linearity
2. σ^2 is constant, independent of x
3. Observations (e 's) are independent of each other
4. For proper statistical inference (CI, P -values), y (e) is normal conditional on x

Verifying some of the assumptions:

O

- In a scattergram the spread of y about the fitted line should be constant as x increases, and y vs. x should appear linear
- Easier to see this with a plot of $\hat{d} = y - \hat{y}$ vs. \hat{y}
- In this plot there are no systematic patterns (no trend in central tendency, no change in spread of points with x)
- Trend in central tendency indicates failure of linearity
- qqnorm plot of d

```
# Fit a model where x and y should have been log transformed
n <- 50
set.seed(2)
res <- rnorm(n, sd=.25)
x <- runif(n)
y <- exp(log(x) + res)
f <- ols(y ~ x)
plot(fitted(f), resid(f))      # Fig. 10.5
# Fit a linear model that should have been quadratic
x <- runif(n, -1, 1)
y <- x ^ 2 + res
f <- ols(y ~ x)
plot(fitted(f), resid(f))
# Fit a correctly assumed linear model
y <- x + res
f <- ols(y ~ x)
plot(fitted(f), resid(f))
# Q-Q plot to check normality of residuals
qqnorm(resid(f)); qqline(resid(f))
```

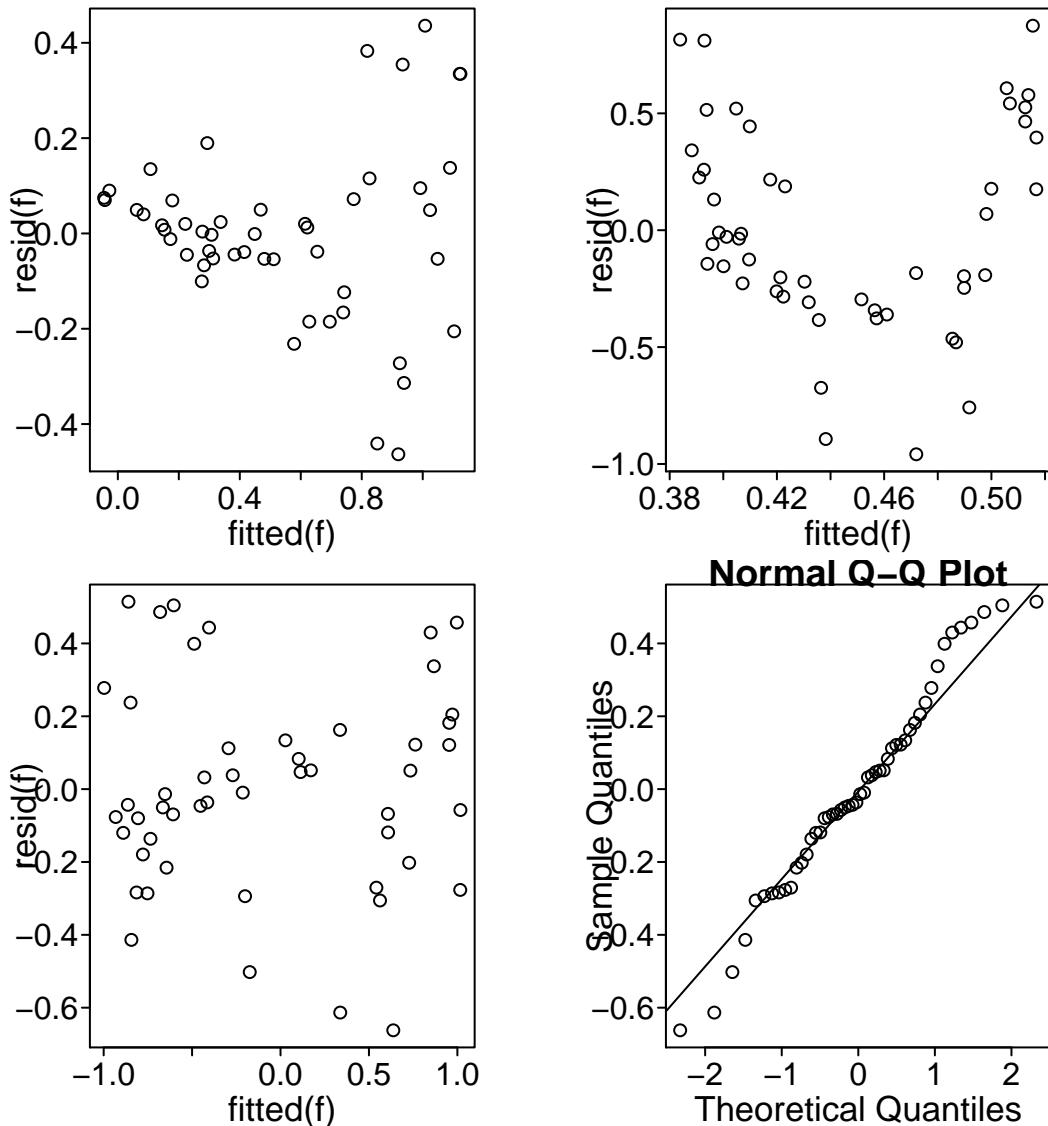


Figure 10.5: Using residuals to check some of the assumptions of the simple linear regression model. Top left panel depicts non-constant σ^2 , which might call for transforming y . Top right panel shows constant variance but the presence of a systemic trend which indicates failure of the linearity assumption. Bottom left panel shows the ideal situation of white noise (no trend, constant variance). Bottom right panel shows a $q-q$ plot that demonstrates approximate normality of residuals, for a sample of size $n = 50$. Horizontal reference lines are at zero, which is by definition the mean of all residuals.

10.5.9

Summary: Useful Equations for Linear Regression

Simple linear regression: one predictor ($p = 1$):

Model: $E(y|x) = \alpha + \beta x$

$E(y)$ = expectation or long-term average of y | = conditional on

Alternate statement of model: $y = \alpha + \beta x + e$, e normal with mean zero for all x , $\text{var}(e) = \sigma^2 = \text{var}(y|x)$

Assumptions:

P

1. Linearity
2. σ^2 is constant, independent of x
3. Observations (e 's) are independent of each other
4. For proper statistical inference (CI, P -values), y (e) is normal conditional on x

Verifying some of the assumptions:

Q

1. In a scattergram the spread of y about the fitted line should be constant as x increases
2. In a residual plot ($d = y - \hat{y}$ vs. x) there are no systematic patterns (no trend in central tendency, no change in spread of points with x)

Sample of size n : $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

R

$$L_{xx} = \sum(x_i - \bar{x})^2 \quad L_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta} = b = \frac{L_{xy}}{L_{xx}} \quad \hat{\alpha} = a = \bar{y} - b\bar{x}$$

$\hat{y} = a + bx = \hat{E}(y|x)$ estimate of $E(y|x)$ = estimate of y

$$SST = \sum(y_i - \bar{y})^2 \quad MST = \frac{SST}{n-1} = s_y^2$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2 \quad MSR = \frac{SSR}{p}$$

$$SSE = \sum(y_i - \hat{y}_i)^2 \quad MSE = \frac{SSE}{n-p-1} = s_{y|x}^2$$

$$\begin{aligned}
SST &= SSR + SSE \quad F = \frac{MSR}{MSE} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \sim F_{p,n-p-1} \\
R^2 &= \frac{SSR}{SST} \quad \frac{SSR}{MSE} \sim \chi_p^2 \\
(p = 1) \quad \widehat{s.e.}(b) &= \frac{s_{y \cdot x}}{\sqrt{L_{xx}}} \\
1 - \alpha \text{ two-sided CI for } \beta & \\
(p = 1) \quad \widehat{s.e.}(\hat{y}) &= s_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}}} \\
1 - \alpha \text{ two-sided CI for } y & \\
(p = 1) \quad \widehat{s.e.}(\hat{E}(y|x)) &= s_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{L_{xx}}} \\
1 - \alpha \text{ two-sided CI for } E(y|x) & \\
& b \pm t_{n-p-1, 1-\alpha/2} \widehat{s.e.}(b) \\
& \hat{y} \pm t_{n-p-1, 1-\alpha/2} \widehat{s.e.}(\hat{y}) \\
& \hat{y} \pm t_{n-p-1, 1-\alpha/2} \widehat{s.e.}(\hat{E}(y|x))
\end{aligned}$$

Multiple linear regression: p predictors, $p > 1$:

Model: $E(y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + e$

Interpretation of β_j : effect on y of increasing x_j by one unit, holding all other x 's constant

Assumptions: same as for $p = 1$ plus no interaction between the x 's (x 's act additively; effect of x_j does not depend on the other x 's).

Verifying some of the assumptions:

S

- When $p = 2$, x_1 is continuous, and x_2 is binary, the pattern of y vs. x_1 , with points identified by x_2 , is two straight, parallel lines
- In a residual plot ($d = y - \hat{y}$ vs. \hat{y}) there are no systematic patterns (no trend in central tendency, no change in spread of points with \hat{y}). The same is true if one plots d vs. any of the x 's.
- Partial residual plots reveal the partial (adjusted) relationship between a chosen x_j and y , controlling for all other $x_i, i \neq j$, without assuming linearity for x_j . In these plots, the following quantities appear on the axes:

y axis: residuals from predicting y from all predictors except x_j

x axis: residuals from predicting x_j from all predictors except x_j (y is ignored)

When $p > 1$, least squares estimates are obtained using more complex formulas. But just as in the case with $p = 1$, all of the coefficient estimates are weighted combinations of the y 's, $\sum w_i y_i$ [when $p = 1$, the w_i for estimating β are $\frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2}$].

Hypothesis tests with $p > 1$:

T

- Overall F test tests $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ vs. the alternative hypothesis that at least one of the β 's $\neq 0$.
- To test whether an individual $\beta_j = 0$ the simplest approach is to compute the t statistic, with $n - p - 1$ d.f.
- Subsets of the β 's can be tested against zero if one knows the standard errors of all of the estimated coefficients and the correlations of each pair of estimates. The formulas are daunting.
- To test whether a subset of the β 's are all zero, a good approach is to compare the model containing all of the predictors associated with the β 's of interest with a sub-model containing only the predictors not being tested (i.e., the predictors being adjusted for). This tests whether the predictors of interest add response information to the predictors being adjusted for. If the goal is to test $H_0 : \beta_1 = \beta_2 = \dots = \beta_q = 0$ regardless of the values of $\beta_{q+1}, \dots, \beta_p$ (i.e., adjusting for x_{q+1}, \dots, x_p), fit the full model with p predictors, computing SSE_{full} or R^2_{full} . Then fit the sub-model omitting x_1, \dots, x_q to obtain $SSE_{reduced}$ or $R^2_{reduced}$. Then compute the partial F statistic

$$F = \frac{(SSE_{reduced} - SSE_{full})/q}{SSE_{full}/(n - p - 1)} = \frac{(R^2_{full} - R^2_{reduced})/q}{(1 - R^2_{full})/(n - p - 1)} \sim F_{q, n-p-1}$$

Note that $SSE_{reduced} - SSE_{full} = SSR_{full} - SSR_{reduced}$.

Notes about distributions:

U

- If $t \sim t_b$, $t \sim \text{normal}$ for large b and $t^2 \sim \chi^2_1$, so $[\frac{b}{\text{s.e.}(b)}]^2 \sim \chi^2_1$
- If $F \sim F_{a,b}$, $a \times F \sim \chi^2_a$ for large b
- If $F \sim F_{1,b}$, $\sqrt{F} \sim t_b$

- If $t \sim t_b$, $t^2 \sim F_{1,b}$
- If $z \sim \text{normal}$, $z^2 \sim \chi_1^2$
- $y \sim D$ means y is distributed as the distribution D
- $y \dot{\sim} D$ means that y is approximately distributed as D for large n
- $\hat{\theta}$ means an estimate of θ

10.6

Proper Transformations and Percentiling



reg-percentiling
V

- All parametric and semi-parametric regression models make assumptions about the shape of the relationship between predictor X and response variable Y
- Many analysts assume linear relationships by default
- Regression splines (piecewise polynomials) are natural nonlinear generalizations
- In epidemiology and public health many practitioners analyze data using percentiling (e.g., of BMI against a random sample of the population)
- This assumes that X affects Y through the population distribution of X (e.g., how many persons have BMI similar to a subject) instead of through physics, physiology, or anatomy
- Also allows definitions to change as the population accommodates
- Example: assume BMI is normal with mean 28 and SD 2 W
- Figure 10.6 upper left panel shows this distribution
- Upper right: percentile of BMI vs. raw BMI
- Lower left: supposed relationship between BMI and disease risk
- Lower right: resulting relationship between BMI percentile and risk

All parametric regression models make assumptions about the form of the relationship between the predictors X and the response variable Y . The typical default assumption is linearity in X vs. some transformation of Y (e.g., log odds, log hazard or in ordinary regression the identity function). Regression splines are one of the best approaches for allowing for smooth, flexible regression function shapes. Splines are described in detail in the *Regression Modeling Strategies* book and course notes.

Some researchers and government agencies get the idea that continuous variables should be modeled through percentiling. This is a rather bizarre way to attempt to account

for shifts in the distribution of the variable by age, race, sex, or geographic location. Percentiling fails to recognize that the way that measurements affect subjects' responses is through physics, physiology, anatomy, and other means. Percentiling in effect states that how a variable affects the outcome for a subject depends on how many other subjects there are like her. When also percentiling variables (such as BMI) over time, the measurement even changes its meaning over time. For example, updating percentiles of BMI each year will keep the fraction of obese members in the population constant even when obesity is truly on the rise.

Putting percentiles into a regression model assumes that the shape of the $X - Y$ relationship is a very strange. As an example, suppose that BMI has a normal distribution with mean 28 and standard deviation 2. The density function for BMI is shown in the upper left panel of Figure 10.6, and the function giving the percentile of BMI as a function of absolute BMI is in the upper right panel.

```
x <- seq(10, 55, length=200)
d <- dnorm(x, mean=28, sd=2)
plot(x, d, type='l', xlab='BMI', ylab='Density')    # Fig. 10.6
pctile <- 100*pnorm(x, mean=28, sd=2)
plot(x, pctile, type='l', xlab='BMI', ylab='BMI Percentile')
risk <- .01 + pmax(x - 25, 0)*.01
plot(x, risk, type='l', xlab='BMI', ylab='Risk')
plot(pctile, risk, type='l', xlab='BMI Percentile', ylab='Risk')
```

Suppose that the true relationship between BMI and the risk of a disease is given in the lower left panel. Then the relationship between BMI percentile and risk must be that shown in the lower right panel. To properly model that shape one must “undo” the percentile function then fit the result with a linear spline. Percentiling creates unrealistic fits and results in more effort being spent if one is to properly model the predictor. X

- Worse still is to group X into quintiles and use a linear model in the quintile numbers
 - assumes a bizarre shape of relationship between X and Y , even if not noticing the discontinuities
- Figure 10.7 depicts quantile numbers vs. mean BMI within each quintile. Outer quintiles:
 - have extremely skewed BMI distributions
 - are too heterogeneous to yield adequate BMI adjustment (residual confounding)

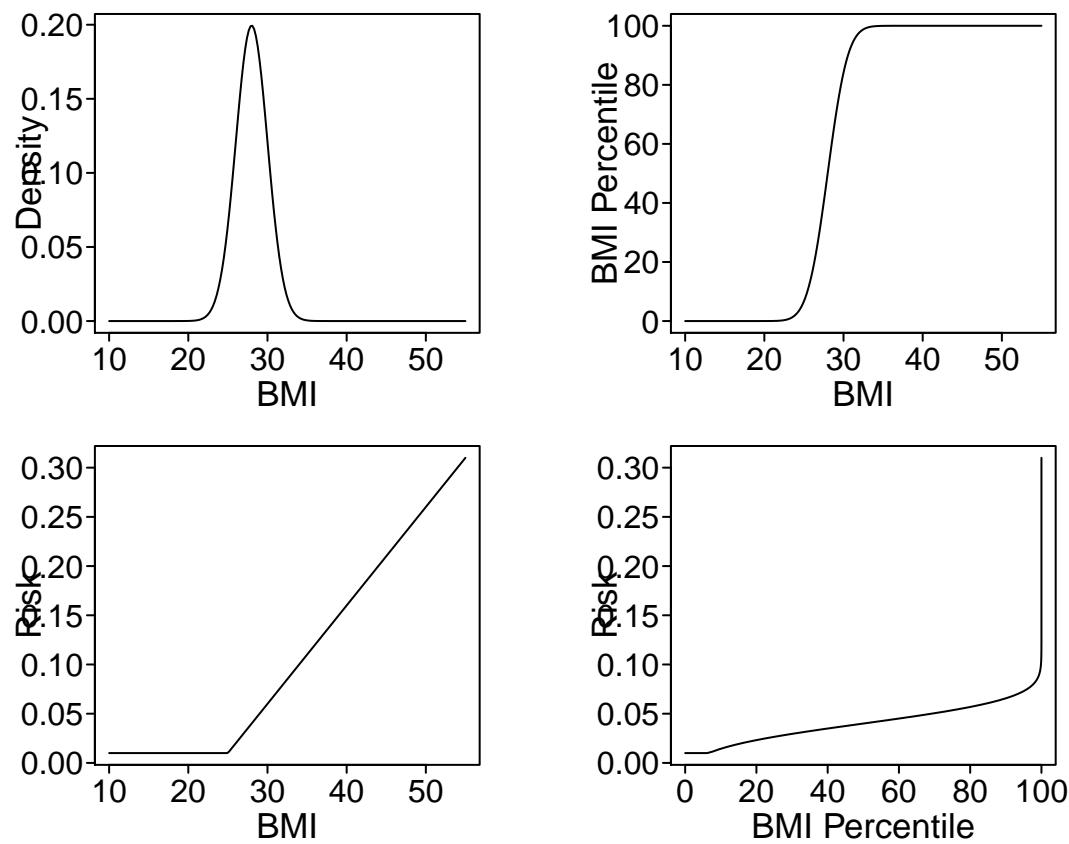


Figure 10.6: Harm of percentiling BMI in a regression model

- Easy to see that the transformation of BMI that yields quintile numbers is discontinuous with variable step widths

In epidemiology a common practice is even more problematic. One often sees smoothly-acting continuous variables such as BMI broken into discontinuous quintile *groups*, the groups numbered from 1 to 5, and a linear regression of correlation fitted to the 1–5 variable (“test for trend”). This is not only hugely wasteful of information and power, but results in significant heterogeneity (especially in the outer quintiles) and assumes a discontinuous effect on outcome that has an exceedingly unusual shape when interpreted on the original BMI scale. Y

Taking the BMI distribution in Figure 10.6 consider what this implies. We draw a random sample of size 500 from the BMI distribution. Figure 10.7 shows the discontinuous relationship between BMI and quintile interval. The location of the mean BMI within BMI quintile is a circle on each horizontal line. One can see the asymmetry of the BMI distribution in the outer quintiles, and that the meaning of inner quantiles is fundamentally different than the meaning of the outer ones because of the narrow range of BMI for inner quantile groups.

```
set.seed(1)
bmi  ← rnorm(500, mean=28, sd=2)
require(Hmisc)
bmiq ← cut2(bmi, g=5)
table(bmiq)
```

bmiq	[22.0,26.5)	[26.5,27.5)	[27.5,28.6)	[28.6,29.8)	[29.8,35.6]
100	100	100	100	100	

```
cuts ← cut2(bmi, g=5, onlycuts=TRUE)
cuts
```

```
[1] 21.98390 26.51345 27.48995 28.55872 29.76222 35.62055
```

```
bmim ← cut2(bmi, g=5, levels.mean=TRUE)
means ← as.numeric(levels(bmim))
plot(c(21, 36), c(1, 5), type='n', xlab='BMI', ylab='Quintile #')    # Fig. 10.7
for(i in 1 : 5) {
  lines(cuts[c(i, i+1)], c(i, i))
  points(means[i], i)
}
```

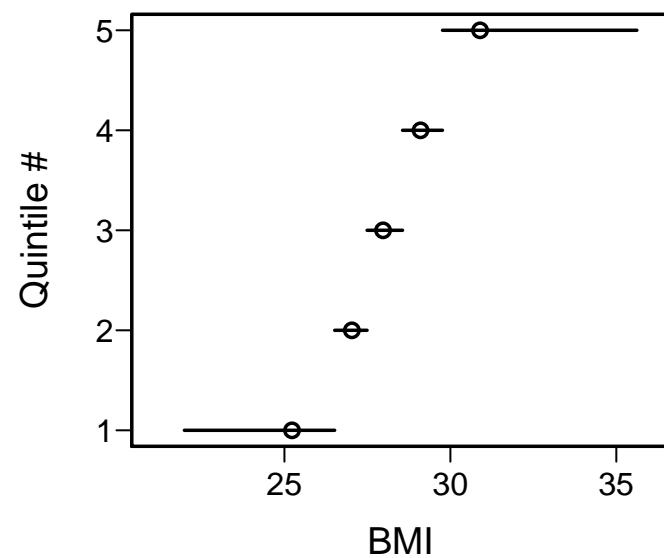


Figure 10.7: What are quintile numbers modeling?

10.7

Multiple Linear Regression

10.7.1

The Model and How Parameters are Estimated

- p independent variables x_1, x_2, \dots, x_p
- Examples: multiple risk factors, treatment plus patient descriptors when adjusting for non-randomized treatment selection in an observational study; a set of controlled or uncontrolled factors in an experimental study; indicators of multiple experimental manipulations performed simultaneously
- Each variable has its own effect (slope) representing *partial effects*: effect of increasing a variable by one unit, holding all others constant
- Initially assume that the different variables act in an additive fashion
- Assume the variables act linearly against y
- Model: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + e$ A
- Or: $E(y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$
- For two x -variables: $y = \alpha + \beta_1 x_1 + \beta_2 x_2$
- Estimated equation: $\hat{y} = a + b_1 x_1 + b_2 x_2$
- Least squares criterion for fitting the model (estimating the parameters):

$$SSE = \sum_{i=1}^n [y_i - (a + b_1 x_{i1} + b_2 x_{i2})]^2$$
- Solve for a, b_1, b_2 to minimize SSE
- When $p > 1$, least squares estimates require complex formulas; still all of the coefficient estimates are weighted combinations of the y 's, $\sum w_i y_i$ ^d.

^dWhen $p = 1$, the w_i for estimating β are $\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$

10.7.2

Interpretation of Parameters

 B

- Regression coefficients are (b) are commonly called *partial regression coefficients*: effects of each variable holding all other variables in the model constant
- Examples of partial effects:

- model containing $x_1 = \text{age (years)}$ and $x_2 = \text{sex (0=male 1=female)}$
Coefficient of age (β_1) is the change in the mean of y for males when age increases by 1 year. It is also the change in y per unit increase in age for females. β_2 is the female minus male mean difference in y for two subjects of the same age.
- $E(y|x_1, x_2) = \alpha + \beta_1 x_1$ for males, $\alpha + \beta_1 x_1 + \beta_2 = (\alpha + \beta_2) + \beta_1 x_1$ for females
[the sex effect is a shift effect or change in y -intercept]
- model with age and systolic blood pressure measured when the study begins
Coefficient of blood pressure is the change in mean y when blood pressure increases by 1mmHg for subjects of the same age

- What is meant by changing a variable?
 - We usually really mean a comparison of two subjects with different blood pressures
 - Or we can envision what would be the expected response had *this* subject's blood pressure been 1mmHg greater at the outset^e
 - We are not speaking of longitudinal changes in a single person's blood pressure
 - We can use subtraction to get the adjusted (partial) effect of a variable, e.g.,

$$E(y|x_1 = a + 1, x_2 = s) - E(y|x_1 = a, x_2 = s) =$$

$$\alpha + \beta_1(a + 1) + \beta_2 s - (\alpha + \beta_1 a + \beta_2 s) = \beta_1$$
- Example: $\hat{y} = 37 + .01 \times \text{weight} + 0.5 \times \text{cigarettes smoked per day}$
 - .01 is the estimate of average increase y across subjects when weight is increased

^eThis setup is the basis for randomized controlled trials and randomized animal experiments. Drug effects may be estimated with between-patient group differences under a statistical model.

D

- by 1lb. if cigarette smoking is unchanged
- 0.5 is the estimate of the average increase in y across subjects per additional cigarette smoked per day if weight does not change
 - 37 is the estimated mean of y for a subject of zero weight who does not smoke
- Comparing regression coefficients:
 - Can't compare directly because of different units of measurement. Coefficients in units of $\frac{y}{x}$.
 - Standardizing by standard deviations: not recommended. Standard deviations are not magic summaries of scale and they give the wrong answer when an x is categorical (e.g., sex).

10.7.3

Example: Estimation of Body Surface Area



DuBois & DuBois developed an equation in log height and log weight in 1916 that is still used^f. We use the main data they used^g.

```
require(rms)
d ← read.csv(textConnection(
'weight,height,bsa
24.2,110.3,8473
64.0,164.3,16720
64.1,178.0,18375
74.1,179.2,19000
93.0,149.7,18592
45.2,171.8,14901
32.7,141.5,11869
6.27,73.2,3699
57.6,164.8,16451
63.0,184.2,17981'))
d ← upData(d, labels=c(weight='Weight', height='Height',
                      bsa='Body Surface Area'),
            units=c(weight='kg', height='cm', bsa='cm^2'), print=FALSE)
d
```

	weight	height	bsa
1	24.20	110.3	8473
2	64.00	164.3	16720
3	64.10	178.0	18375

^fDuBois D, DuBois EF: A formula to estimate the approximate surface area if height and weight be known. *Arch Int Medicine* 17(6):863-71, 1916.

^gA Stata data file `dubois.dta` is available [here](#).

4	74.10	179.2	19000
5	93.00	149.7	18592
6	45.20	171.8	14901
7	32.70	141.5	11869
8	6.27	73.2	3699
9	57.60	164.8	16451
10	63.00	184.2	17981

```
# Create Stata file
getRs('r2stata.r', put='source')
dubois <- d
r2stata(dubois)
# Exclude subject measured using adhesive plaster method
d <- d[-7, ]
```

Fit a multiple regression model in the logs of all 3 variables

E

```
dd <- datadist(d); options(datadist='dd')
f <- ols(log10(bsa) ~ log10(weight) + log10(height), data=d)
f
```

Linear Regression Model

```
ols(formula = log10(bsa) ~ log10(weight) + log10(height), data = d)
```

	Model Likelihood Ratio Test		Discrimination Indexes	
Obs 9	LR χ^2	66.23	R^2	0.999
σ 0.0069	d.f.	2	R_{adj}^2	0.999
d.f. 6	Pr(> χ^2)	0.0000	g	0.226

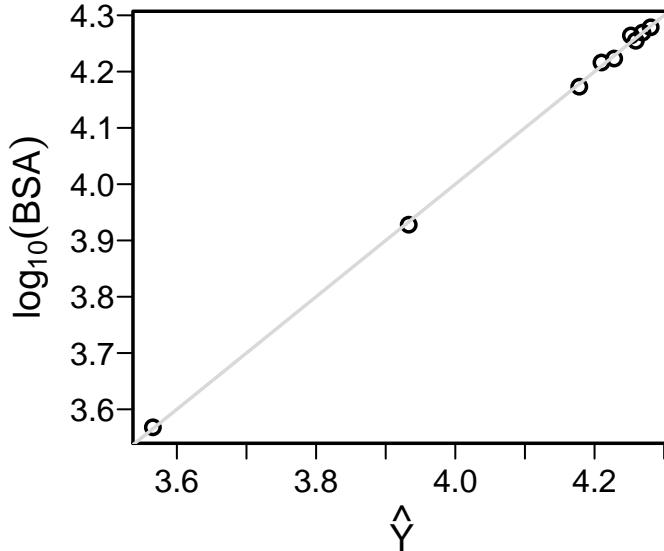
Residuals				
Min	1Q	Median	3Q	Max
-0.005031	-0.004851	-0.001908	0.002541	0.01198

	$\hat{\beta}$	S.E.	t	Pr(> t)
Intercept	1.9607	0.0808	24.26	<0.0001
weight	0.4198	0.0184	22.77	<0.0001
height	0.6812	0.0499	13.64	<0.0001

DuBois & DuBois derived the equation $\log(\text{bsa}) = 1.8564 + 0.426 \log(\text{weight}) + 0.725 \log(\text{height})$

Plot predicted vs. observed

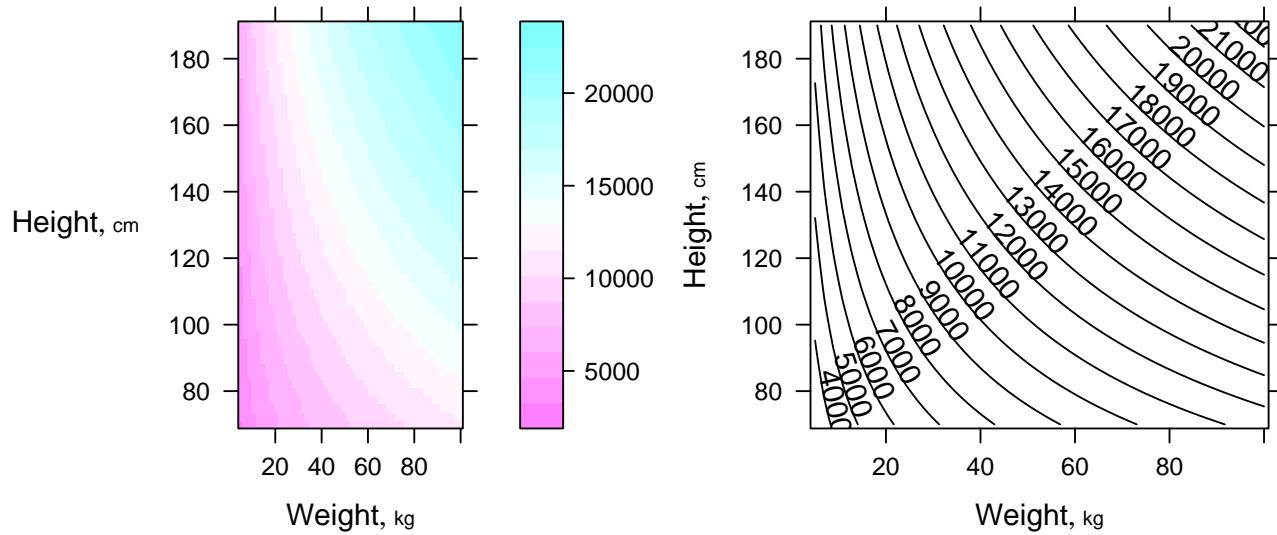
```
plot(fitted(f), log10(d$bsa), xlab=expression(hat(Y)),
      ylab=expression(log[10](BSA))); abline(a=0, b=1, col=gray(.85))
```



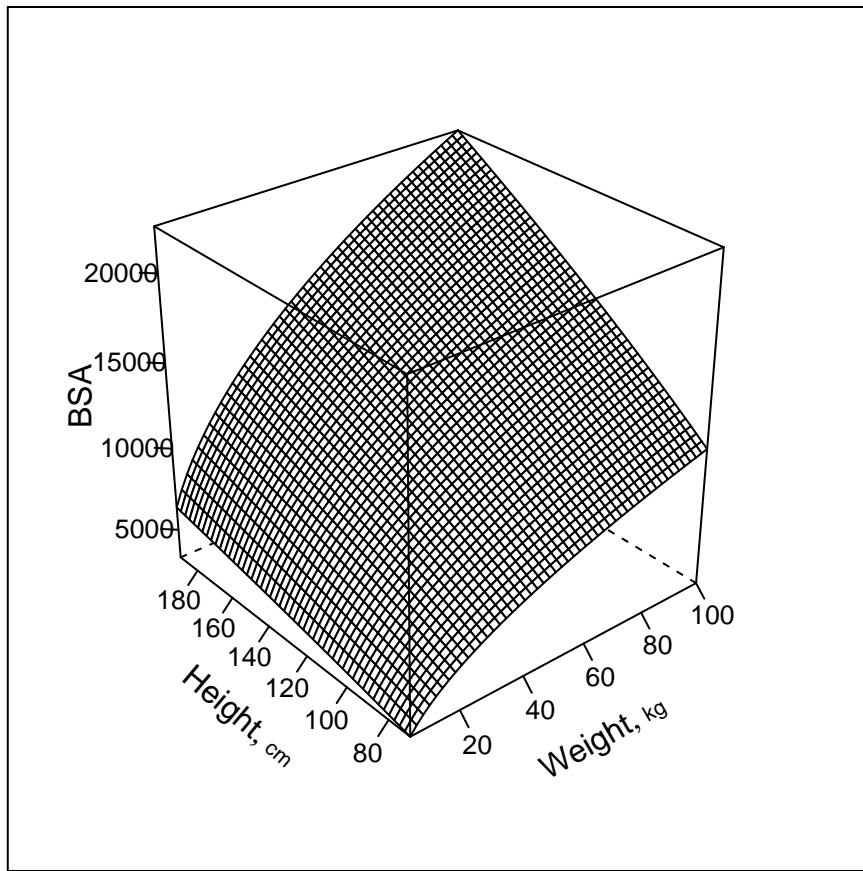
Get 3 types of plots to show fitted model

```
p <- Predict(f, weight=seq(5, 100, length=50),
              height=seq(70, 190, length=50), fun=function(z) 10 ^ z)
p1 <- bplot(p)
p2 <- bplot(p, lfun=contourplot, cuts=20)
arrGrob(p1, p2, ncol=2)
```

F



```
bplot(p, lfun=wireframe, zlab='BSA')
```



Note: this plot would be a plane if all 3 variables were plotted on the scale fitted in the regression (\log_{10}).

10.7.4

What are Degrees of Freedom



For a model : the total number of parameters not counting intercept(s)

For a hypothesis test : the number of parameters that are hypothesized to equal specified constants. The constants specified are usually zeros (for *null* hypotheses) but this is not always the case. Some tests involve combinations of multiple parameters but test this combination against a single constant; the d.f. in this case is still one. Example: $H_0 : \beta_3 = \beta_4$ is the same as $H_0 : \beta_3 - \beta_4 = 0$ and is a 1 d.f. test because it tests one parameter ($\beta_3 - \beta_4$) against a constant (0).

These are **numerator d.f.** in the sense of the *F*-test in multiple linear regression. The

F -test also entails a second kind of d.f., the **denominator** or **error** d.f., $n - p - 1$, where p is the number of parameters aside from the intercept. The error d.f. is the denominator of the estimator for σ^2 that is used to unbias the estimator, penalizing for having estimated $p + 1$ parameters by minimizing the sum of squared errors used to estimate σ^2 itself. You can think of the error d.f. as the sample size penalized for the number of parameters estimated, or as a measure of the information base used to fit the model.

Other ways to express the d.f. for a hypothesis are:

H

- The number of opportunities you give associations to be present (relationships with Y to be non-flat)
- The number of restrictions you place on parameters to make the null hypothesis of no association (flat relationships) hold

10.7.5

Hypothesis Testing

Testing Total Association (Global Null Hypotheses)

reg-mult-h0

- ANOVA table is same as before for general p
- $F_{p,n-p-1}$ tests $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$
- This is a test of *total association*, i.e., a test of whether *any* of the predictors is associated with y
- To assess total association we accumulate partial effects of all variables in the model *even though* we are testing if *any* of the partial effects is nonzero
- H_a : at least one of the β 's is non-zero. **Note:** This does not mean that all of the x variables are associated with y .
- Weight and smoking example: H_0 tests the null hypothesis that neither weight nor smoking is associated with y . H_a is that at least one of the two variables is

associated with y . The other may or may not have a non-zero β .

- Test of total association does not test whether cigarette smoking is related to y holding weight constant.
- SSR can be called the model SS

Testing Partial Effects



10e-multiple-partial

J

- $H_0 : \beta_1 = 0$ is a test for the effect of x_1 on y holding x_2 and any other x 's constant
- Note that β_2 is *not* part of the null or alternative hypothesis; we assume that we have adjusted for *whatever* effect x_2 has, *if any*
- One way to test β_1 is to use a t -test: $t_{n-p-1} = \frac{b_1}{\widehat{s.e.}(b_1)}$
- In multiple regression it is difficult to compute standard errors so we use a computer
- These standard errors, like the one-variable case, decrease when
 - $n \uparrow$
 - variance of the variable being tested \uparrow
 - σ^2 (residual y -variance) \downarrow
- Another way to get partial tests: the F test
 - Gives identical 2-tailed P -value to t test when one x being tested
 $t^2 \equiv \text{partial } F$
 - Allows testing for > 1 variable
 - Example: is either systolic or diastolic blood pressure (or both) associated with the time until a stroke, holding weight constant
- To get a partial F define partial SS

K

- Partial SS is the change in SS when the variables **being tested** are dropped from the model and the model is re-fitted
- A general principle in regression models: a set of variables can be tested for their combined partial effects by removing that set of variables from the model and measuring the harm ($\uparrow SSE$) done to the model L
- Let *full* refer to computed values from the full model including all variables; *reduced* denotes a reduced model containing only the adjustment variables and not the variables being tested
- Dropping variables $\uparrow SSE, \downarrow SSR$ unless the dropped variables had exactly zero slope estimates in the full model (which never happens)
- $SSE_{reduced} - SSE_{full} = SSR_{full} - SSR_{reduced}$
Numerator of F test can use either SSE or SSR
- Form of partial F -test: change in SS when dropping the variables of interest divided by change in d.f., then divided by MSE ;
 MSE is chosen as that which best estimates σ^2 , namely the MSE from the full model
- Full model has p slopes; suppose we want to test q of the slopes M

$$\begin{aligned} F_{q,n-p-1} &= \frac{(SSE_{reduced} - SSE_{full})/q}{MSE} \\ &= \frac{(SSR_{full} - SSR_{reduced})/q}{MSE} \end{aligned}$$

10.7.6

Assessing Goodness of Fit

Assumptions:

- Linearity of each predictor against y holding others constant
- σ^2 is constant, independent of x

- Observations (e 's) are independent of each other
- For proper statistical inference (CI, P -values), y (e) is normal conditional on x
- x 's act additively; effect of x_j does not depend on the other x 's (But note that the x 's may be correlated with each other without affecting what we are doing.)

Verifying some of the assumptions:

o

1. When $p = 2$, x_1 is continuous, and x_2 is binary, the pattern of y vs. x_1 , with points identified by x_2 , is two straight, parallel lines. β_2 is the slope of y vs. x_2 holding x_1 constant, which is just the difference in means for $x_2 = 1$ vs. $x_2 = 0$ as $\Delta x_2 = 1$ in this simple case.

```
# Generate 25 observations for each group, with true beta1=.05, true beta2=3
d <- expand.grid(x1=1:25, x2=c(0, 1))
set.seed(3)
d$y <- with(d, .2*x1 + 3*x2 + rnorm(50, sd=.5))
with(d, plot(x1, y, xlab=expression(x[1]), ylab=expression(y)))
abline(a=0, b=.2)      # Fig. 10.8
abline(a=3, b=.2)
text(13, 1.3, expression(y==alpha + beta[1]*x[1]), srt=24, cex=1.3)
text(13, 7.1, expression(y==alpha + beta[1]*x[1] + beta[2])), srt=24, cex=1.3)
```

2. In a residual plot ($d = y - \hat{y}$ vs. \hat{y}) there are no systematic patterns (no trend in central tendency, no change in spread of points with \hat{y}). The same is true if one plots d vs. any of the x 's (these are more stringent assessments). If x_2 is binary box plots of d stratified by x_2 are effective.
3. Partial residual plots reveal the partial (adjusted) relationship between a chosen x_j and y , controlling for all other $x_i, i \neq j$, without assuming linearity for x_j . In these plots, the following quantities appear on the axes:

y axis: residuals from predicting y from all predictors except x_j

x axis: residuals from predicting x_j from all predictors except x_j (y is ignored)

Partial residual plots ask how does what we can't predict about y without knowing x_j depend on what we can't predict about x_j from the other x 's.

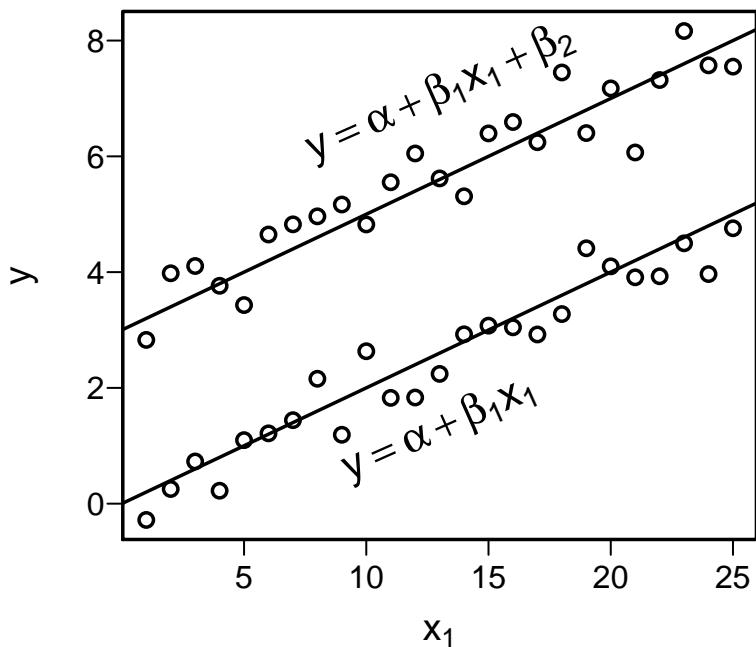


Figure 10.8: Data satisfying all the assumptions of simple multiple linear regression in two predictors. Note equal spread of points around the population regression lines for the $x_2 = 1$ and $x_2 = 0$ groups (upper and lower lines, respectively) and the equal spread across x_1 . The $x_2 = 1$ group has a new intercept, $\alpha + \beta_2$, as the x_2 effect is β_2 . On the y axis you can clearly see the difference between the two true population regression lines is $\beta_2 = 3$.

10.8

Multiple Regression with a Binary Predictor

10.8.1

Indicator Variable for Two-Level Categorical Predictors

reg-xbinary

 P

- Categories of predictor: A, B (for example)
- First category = reference cell, gets a zero
- Second category gets a 1.0
- Formal definition of indicator (dummy) variable: $x = [category = B]$
 $[w] = 1$ if w is true, 0 otherwise
- $\alpha + \beta x = \alpha + \beta[category = B] =$
 α for category A subjects
 $\alpha + \beta$ for category B subjects
 β = mean difference ($B - A$)

10.8.2

Two-Sample t -test vs. Simple Linear Regression

Q

- They are equivalent in every sense:
 - P -value
 - Estimates and C.L.s after rephrasing the model
 - Assumptions (equal variance assumption of two groups in t -test is the same as constant variance of $y|x$ for every x)
- $a = \bar{Y}_A$
 $b = \bar{Y}_B - \bar{Y}_A$

- $\widehat{s.e.}(b) = \widehat{s.e.}(\bar{Y}_B - \bar{Y}_A)$

10.8.3

Analysis of Covariance



reg-anova
R

- Multiple regression can extend the *t*-test
 - More than 2 groups (multiple indicator variables can do multiple-group ANOVA)
 - Allow for categorical or continuous adjustment variables (covariates, covari-
ables)
- Example: lead exposure and neuro-psychological function (Rosner)
- Model: $MAXFWT = \alpha + \beta_1 age + \beta_2 sex + e$
- Rosner coded $sex = 1, 2$ for male, female
Does not affect interpretation of β_2 but makes interpretation of α more tricky
(mean $MAXFWT$ when $age = 0$ and $sex = 0$ which is impossible by this coding.)
- Better coding would have been $sex = 0, 1$ for male, female
 - α = mean $MAXFWT$ for a zero year-old male
 - β_1 = increase in mean $MAXFWT$ per 1-year increase in age
 - β_2 = mean $MAXFWT$ for females minus mean $MAXFWT$ for males, hold-
ing age constant
- Suppose that we define an (arbitrary) exposure variable to mean that the lead dose S
 $\geq 40\text{mg}/100\text{ml}$ in either 1972 **or** 1973
- Model: $MAXFWT = \alpha + \beta_1 exposure + \beta_2 age + \beta_3 sex + e$
 $exposure = \text{TRUE } (1) \text{ for exposed, FALSE } (0) \text{ for unexposed}$
- β_1 = mean $MAXFWT$ for exposed minus mean for unexposed, holding age and sex constant

10.9

The Correlation Coefficient Revisited



reg-corr
T

Pearson product-moment linear correlation coefficient:

$$\begin{aligned} r &= \frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \\ &= \frac{s_{xy}}{s_x s_y} \\ &= b \sqrt{\frac{L_{xx}}{L_{yy}}} \end{aligned}$$

U

- r is unitless
- r estimates the population correlation coefficient ρ (not to be confused with Spearman ρ rank correlation coefficient)
- $-1 \leq r \leq 1$
- $r = -1$: perfect negative correlation
- $r = 1$: perfect positive correlation
- $r = 0$: no correlation (no association)
- t – test for r is identical to t -test for b
- r^2 is the proportion of variation in y explained by conditioning on x
- $(n - 2) \frac{r^2}{1-r^2} = F_{1,n-2} = \frac{MSR}{MSE}$
- For multiple regression in general we use R^2 to denote the fraction of variation in y explained jointly by all the x 's (variation in y explained by the whole model) V
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1$ minus fraction of unexplained variation
- R^2 is called the *coefficient of determination*

- R^2 is between 0 and 1
 - 0 when $\hat{y}_i = \bar{y}$ for all i ; $SSE = SST$
 - 1 when $\hat{y}_i = y_i$ for all i ; $SSE=0$
- $R^2 \equiv r^2$ in the one-predictor case

10.10

Using Regression for ANOVA

10.10.1

Indicator Variables

Lead Exposure Group (Rosner lead dataset):

control : normal in both 1972 and 1973

currently exposed : elevated serum lead level in 1973, normal in 1972

previously exposed : elevated lead in 1972, normal in 1973

NOTE: This is not a very satisfactory way to analyze the two years' worth of lead exposure data, as we do not expect a discontinuous relationship between lead levels and neurological function. A continuous analysis was done in Chapter 9.

ABD18.1
reg-anova
W



W

X

- Requires two indicator (dummy) variables (and 2 d.f.) to perfectly describe 3 categories
- $x_1 = [\text{currently exposed}]$
- $x_2 = [\text{previously exposed}]$
- Reference cell is control
- lead dataset group variable is set up this way already
- Model:

Y

$$\begin{aligned}
 E(y|\text{exposure}) &= \alpha + \beta_1 x_1 + \beta_2 x_2 \\
 &= \alpha, \text{controls} \\
 &= \alpha + \beta_1, \text{currently exposed} \\
 &= \alpha + \beta_2, \text{previously exposed}
 \end{aligned}$$

α : mean maxfwt for controls

β_1 : mean maxfwt for currently exposed minus mean for controls

β_2 : mean maxfwt for previously exposed minus mean for controls

$\beta_2 - \beta_1$: mean for previously exposed minus mean for currently exposed

```
getHdata(lead)
dd <- datadist(lead); options(datadist='dd')
f <- ols(maxfwt ~ group, data=lead)
f
```

Linear Regression Model

```
ols(formula = maxfwt ~ group, data = lead)
```

Frequencies of Missing Values Due to Each Variable

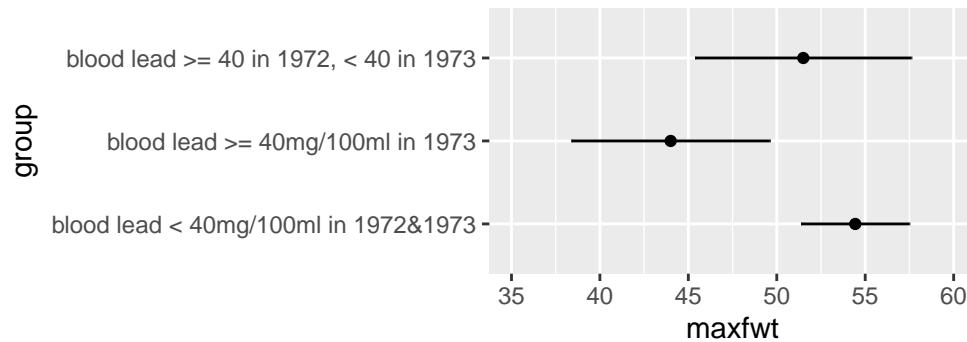
maxfwt	group
25	0

	Model Likelihood		Discrimination	
	Ratio Test		Indexes	
Obs	99	LR χ^2	10.33	R^2 0.099
σ	12.3127	d.f.	2	R^2_{adj} 0.080
d.f.	96	Pr(> χ^2)	0.0057	g 3.706

	Residuals				
	Min	1Q	Median	3Q	Max
	-41.44	-5.75	1.554e-15	7.531	31.5

	$\hat{\beta}$	S.E.	t	Pr(> t)
Intercept	54.4375	1.5391	35.37	<0.0001
group=blood lead \geq 40mg/100ml in 1973	-10.4375	3.2168	-3.24	0.0016
group=blood lead \geq 40 in 1972, < 40 in 1973	-2.9375	3.4415	-0.85	0.3955

```
ggplot(Predict(f))
```



```
options(prType='plain')
summary(f)
```

Effects		Response : maxfwt				
Factor						
group	- blood lead >= 40mg/100ml in 1973:	blood lead < 40mg/100ml in 1972&1973				
group	- blood lead >= 40 in 1972, < 40 in 1973:	blood lead < 40mg/100ml in 1972&1973				
Low	High	Diff. Effect	S.E.	Lower 0.95	Upper 0.95	
1	2	NA	-10.4380	3.2168	-16.8230	-4.0522
1	3	NA	-2.9375	3.4415	-9.7688	3.8938

```
options(prType='latex')
```

A

- In general requires $k - 1$ dummies to describe k categories
- For testing or prediction, choice of reference cell is irrelevant
- Does matter for interpreting individual coefficients
- Modern statistical programs automatically generate indicator variables from categorical or character predictors^h
- In R never generate indicator variables yourself; just provide a factor or character predictor.

^hIn R indicators are generated automatically any time a factor or category variable is in the model.

10.10.2

Obtaining ANOVA with Multiple Regression

 B

- Estimate α, β_j using standard least squares
- F -test for overall regression is exactly F for ANOVA
- In ANOVA, SSR is call sum of squares between treatments
- SSE is called sum of squares within treatments
- Don't need to learn formulas specifically for ANOVA

10.10.3

One-Way Analysis of Covariance

C

- Just add other variables (covariates) to the model
- Example: predictors age and treatment
age is the covariate (adjustment variable)
- Global F test tests the global null hypothesis that neither age nor treatment is associated with response
- To test the adjusted treatment effect, use the partial F test for treatment based on the partial SS for treatment adjusted for age
- If treatment has only two categories, the partial t -test is an easier way to get the age-adjusted treatment test

```
fa <- ols(maxfwt ~ age + group, data=lead)
fa
```

Linear Regression Model

```
ols(formula = maxfwt ~ age + group, data = lead)
```

Frequencies of Missing Values Due to Each Variable

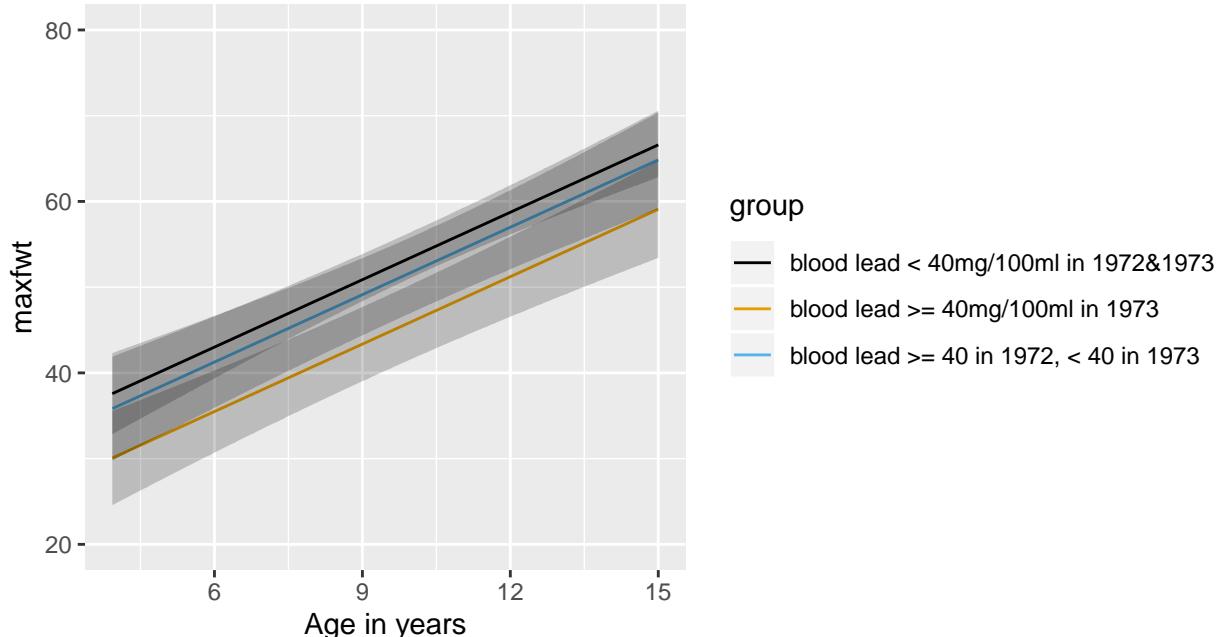
```
maxfwt    age   group
      25      0      0
```

	Model Likelihood Ratio Test		Discrimination Indexes	
Obs	99	LR χ^2	62.98	R^2 0.471
σ	9.4872	d.f.	3	R^2_{adj} 0.454
d.f.	95	Pr(> χ^2)	0.0000	g 10.145

Residuals					
	Min	1Q	Median	3Q	Max
	-33.5	-5.125	0.9098	5.371	33

	$\hat{\beta}$	S.E.	t	Pr(> t)
Intercept	27.2810	3.5303	7.73	<0.0001
age	2.6211	0.3209	8.17	<0.0001
group=blood lead \geq 40mg/100ml in 1973	-7.5148	2.5043	-3.00	0.0034
group=blood lead \geq 40 in 1972, < 40 in 1973	-1.7464	2.6557	-0.66	0.5124

```
ggplot(Predict(fa, age, group))
```



```
options(prType='plain')
summary(fa)
```

Effects

Response : maxfwt

```

Factor
age
group - blood lead >= 40mg/100ml in 1973:blood lead < 40mg/100ml in 1972&1973
group - blood lead >= 40 in 1972, < 40 in 1973:blood lead < 40mg/100ml in 1972&1973
Low    High   Diff. Effect S.E. Lower 0.95 Upper 0.95
6.1667 12.021 5.8542 15.3440 1.8789 11.6140 19.0740
1.0000  2.000      NA -7.5148 2.5043 -12.4860 -2.5431
1.0000  3.000      NA -1.7464 2.6557 -7.0187  3.5259

```

```
options(prType='latex')
```

```
anova(fa)
```

Analysis of Variance for maxfwt

	d.f.	Partial SS	MS	F	P
age	1	6003.1719	6003.17189	66.70	<0.0001
group	2	810.4561	405.22806	4.50	0.0135
REGRESSION	3	7603.2603	2534.42009	28.16	<0.0001
ERROR	95	8550.5781	90.00609		

```
anova(f) # reduced model (without age)
```

Analysis of Variance for maxfwt

	d.f.	Partial SS	MS	F	P
group	2	1600.088	800.0442	5.28	0.0067
REGRESSION	2	1600.088	800.0442	5.28	0.0067
ERROR	96	14553.750	151.6016		

Subtract SSR or SSE from these two models to get the treatment effect with 2 d.f.

10.10.4

Continuous Analysis of Lead Exposure

See 9.4.

10.10.5

Two-Way ANOVA



- Two categorical variables as predictors

- Each variable is expanded into indicator variables
- One of the predictor variables may not be time or episode within subject; two-way ANOVA is often misused for analyzing repeated measurements within subject
- Example: 3 diet groups (NOR, SV, LV) and 2 sex groups
- $E(y|diet, sex) = \alpha + \beta_1[SV] + \beta_2[LV] + \beta_3[male]$
- Assumes effects of diet and sex are additive (separable) and not synergistic
- $\beta_1 = SV - NOR$ mean difference holding sex constant
 $\beta_3 = male - female$ effect holding diet constant
- Test of diet effect controlling for sex effect:
 $H_0 : \beta_1 = \beta_2 = 0$
 $H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$
- This is a 2 d.f. partial F -test, best obtained by taking difference in SS between this full model and a model that excludes all diet terms.
- Test for significant difference in mean y for males vs. females, controlling for diet:
 $H_0 : \beta_3 = 0$
- For a model that has m categorical predictors (only), none of which interact, with numbers of categories given by k_1, k_2, \dots, k_m , the total numerator regression d.f. is $\sum_{i=1}^m (k_i - 1)$

10.10.6

Two-way ANOVA and Interaction

Example: sex (F,M) and treatment (A,B)

Reference cells: F, A

Model:

$$\begin{aligned} E(y|sex, treatment) &= \alpha + \beta_1[sex = M] \\ &\quad + \beta_2[treatment = B] + \beta_3[sex = M \cap treatment = B] \end{aligned}$$

Note that $[sex = M \cap treatment = B] = [sex = M] \times [treatment = B]$. E

α : mean y for female on treatment A (all variables at reference values)

β_1 : mean y for males minus mean for females, both on treatment $A =$ sex effect holding treatment constant at A

β_2 : mean for female subjects on treatment B minus mean for females on treatment $A =$ treatment effect holding sex constant at *female*

β_3 : $B - A$ treatment difference for males minus $B - A$ treatment difference for females
Same as $M - F$ difference for treatment B minus $M - F$ difference for treatment A

In this setting think of interaction as a “double difference”. To understand the parameters: F

Group	$E(y Group)$
F A	α
M A	$\alpha + \beta_1$
F B	$\alpha + \beta_2$
M B	$\alpha + \beta_1 + \beta_2 + \beta_3$

Thus $MB - MA - [FB - FA] = \beta_2 + \beta_3 - \beta_2 = \beta_3$.

Heterogeneity of Treatment Effect

Consider a Cox proportional hazards model for time until first major cardiovascular event. The application is targeted pharmacogenomics in acute coronary syndrome (the CURE study⁷⁸). G



reg-haz

- Subgroup analysis is virtually worthless for learning about differential treatment effects

- Instead a proper assessment of interaction must be used, with liberal adjustment for main effects
- An interaction effect is a double difference; for logistic and Cox models it is the ratio of ratios
- Interactions are harder to assess than main effects (wider confidence intervals, lower power)
- Carriers for loss-of-function CYP2C19 alleles: reduced conversion of clopidogrel to active metabolite H
- Suggested that clop. less effective in reducing CV death, MI, stroke
- 12,562 (clop. HR 0.8); 5059 genotyped (clop. HR 0.7)

	Carrier	Non-Carrier
HR	0.69 (0.49, 0.98)	0.72 (0.59, 0.87)
Ratio of HRs	0.96 (0.64, 1.43)	($P = 0.8$)

|

- In the publication the needed ratio of hazard ratios was nowhere to be found
- C.L. for ratio of hazard ratios shows that CYP2C19 variants may plausibly be associated with a huge benefit or huge harm
- Point estimate is in the wrong direction
- Epidemiologic evidence points to a dominant effect of smoking in this setting
 - Significant interaction between smoking status and clop. effect
 - Lack of evidence that clop. is effective in non-smokers
 - Gurbel, Nolin, Tantry, JAMA 2012;307:2495

10.10.7

Interaction Between Categorical and Continuous Variables



regression

This is how one allows the slope of a predictor to vary by categories of another variable.

Example: separate slope for males and females:

$$\begin{aligned}
 E(y|x) &= \alpha + \beta_1 age + \beta_2 [sex = m] \\
 &\quad + \beta_3 age \times [sex = m] \\
 E(y|age, sex = f) &= \alpha + \beta_1 age \\
 E(y|age, sex = m) &= \alpha + \beta_1 age + \beta_2 + \beta_3 age \\
 &= (\alpha + \beta_2) + (\beta_1 + \beta_3) age
 \end{aligned}$$

J

α : mean y for zero year-old female

β_1 : slope of age for females

β_2 : mean y for males minus mean y for females, for zero year-olds

β_3 : increment in slope in going from females to males

10.11

Internal vs. External Model Validation

External validation or validation on a holdout sample, when the predictive method was developed using feature selection, model selection, or machine learning, produces a non-unique *example* validation of a non-unique *example* model.

FE Harrell, 2015

reg-val
blog
K

- Many researchers assume that “external” validation of model predictions is the only way to have confidence in predictions
- External validation may take years and may be low precision (wide confidence intervals for accuracy estimates)
- Splitting one data sequence to create a holdout sample is *internal* validation, not *external*, and resampling procedures using all available data are almost always better
- External validation by splitting in time or place loses opportunity for modeling secular and geographic trends, and often results in failure to validate when in fact there are interesting group differences or time trends that could have easily been modeled
- One should use all data available at analysis time
- External validation is left for newly collected data not available at publication time
- Rigorous internal validation should be done first
 - “optimism” bootstrap generally has lowest mean squared error of accuracy estimates
 - bootstrap estimates the likely future performance of model developed on whole dataset
 - all analytical steps using Y must be repeated for each of approx. 300-400 bootstrap repetitions
 - when empirical feature selection is part of the process, the bootstrap reveals the true volatility in the list of selected predictors

L

- Many data splitting or external validations are unreliable (example: volatility of splitting 17,000 ICU patients with high mortality, resulting in multiple splits giving different models and different performance in holdout samples) M

There are subtleties in what holdout sample validation actually means, depending on how the predictive model is fitted: N

- When the model's form is fully pre-specified, the external validation validates that model and its estimated coefficients
- When the model is derived using feature selection or machine learning methods, the holdout sample validation is not "honest" in a certain sense:
 - data are incapable of informing the researcher what the "right" predictors and the "right" model are
 - the process doesn't recognize that the model being validated is nothing more than an "example" model
- Resampling for rigorous internal validation validates the process used to derive the "final" model
 - as a byproduct estimates the likely future performance of that model
 - while reporting volatility instead of hiding it

Model validation is a very important and complex topic that is covered in detail in the two books mentioned below. One of the most difficult to understand elements of validation is what is, and when to use, external validation. Some researchers have published predictive tools with no validation at all while other researchers falsely believe that "external" validation is the only valid approach to having confidence in predictions. Frank Harrell (author of *Regression Modeling Strategies*) and Ewout Steyerberg (author of *Clinical Prediction Models*) have written the text below in an attempt to illuminate several issues.

There is much controversy about the need for, definition of, and timing of external validation. A prognostic model should be valid outside the specifics of the sample where the model is developed. Ideally, a model is shown to predict accurately across a wide range of settings (Justice et al, *Ann Int Med* 1999). Evidence of such external validity requires evaluation by different research groups and may take several years. Researchers frequently make the mistake of labeling data splitting from a single sequence of patients as external validation when in fact this is a particularly low-precision form of internal

validation better done using resampling (see below). On the other hand, external validation carried out by splitting in time (temporal validation) or by place, is better replaced by considering interactions in the full dataset. For example, if a model developed on Canadians is found to be poorly calibrated for Germans, it is far better to develop an international model with country as one of the predictors. This implies that a researcher with access to data is always better off to analyze and publish a model developed on the full set. That leaves external validation using (1) newly collected data, not available at the time of development; and (2) other investigators, at other sites, having access to other data. (2) has been promoted by Justice as the strongest form of external validation. This phase is only relevant once internal validity has been shown for the developed model. But again, if such data were available at analysis time, those data are too valuable not to use in model development.

Even in the small subset of studies comprising truly external validations, it is a common misconception that the validation statistics are precise. Many if not most external validations are unreliable due to instability in the estimate of predictive accuracy. This instability comes from two sources: the size of the validation sample, and the constitution of the validation sample. The former is easy to envision, while the latter is more subtle. In one example, Frank Harrell analyzed 17,000 ICU patients with $\frac{1}{3}$ of patients dying, splitting the dataset into two halves - a training sample and a validation sample. He found that the validation c-index (ROC area) changed substantially when the 17,000 patients were re-allocated at random into a new training and test sample and the entire process repeated. Thus it can take quite a large external sample to yield reliable estimates and to "beat" strong internal validation using resampling. Thus we feel there is great utility in using strong internal validation.

At the time of model development, researchers should focus on showing internal validity of the model they propose, i.e. validity of the model for the setting that they consider. Estimates of model performance are usually optimistic. The optimism can efficiently be quantified by a resampling procedure called the bootstrap, and the optimism can be subtracted out to obtain an unbiased estimate of future performance of the model on the same types of patients. The bootstrap, which enjoys a strong reputation in data analysis, entails drawing patients from the development sample with replacement. It allows one to estimate the likely future performance of a predictive model without waiting for new data to perform a external validation study. It is important that the bootstrap model validation be done rigorously. This means that all analytical steps that use the outcome variable are repeated in each bootstrap sample. In this way, the proper price is paid for any statistical assessments to determine the final model, such as choosing variables and estimating regression coefficients. When the resampling allows models and coefficients to disagree with themselves over hundreds of resamples, the proper price is paid for "data dredging", so that clinical utility (useful predictive discrimination) is not claimed for what is in fact overfitting (fitting "noise"). The bootstrap makes optimal use of the available data: it uses all data to develop the model and all data to internally validate the model, detecting any overfitting. One can call properly penalized bootstrapping rigorous or strong internal validation.

To properly design or interpret a predictive model validation it is important to take into account how the predictions were formed. There are two overall classes of predictive approaches:

- formulating a pre-specified statistical model based on understanding of the literature and subject matter and using the data to fit the parameters of the model but not to select the *form* of the model or the features to use as predictors
- using the data to screen candidate predictive features or to choose the form the predictions take. Examples of this class include univariable feature screening, stepwise regression, and machine learning algorithms.

External holdout sample validation may seem to be appropriate for the second class but actually it is not “honest” in the following sense. The data are incapable of informing the researcher of what are the “right” predictors. This is even more true when candidate predictors are correlated, creating competition among predictors and arbitrary selections from these co-linear candidates. A predictive rule derived from the second class of approaches is merely an *example* of a predictive model. The only way for an analyst to understand this point is to use resampling (bootstrap or cross-validation) whereby predictive models (or machine learning structures) are repeatedly derived from scratch and the volatility (and difficulty of the task) are exposed.

So what goes in in the training and validation processes depends on the class of predictive methods used:

Q

- When the model is pre-specified except for the regression coefficients that need to be estimated, rigorous resampling validation validates the fit of *the* model and so does holdout sample validation.
- When the model structure was not pre-specified but model/feature selection was done, resampling validation validates the *process* used to derive the “final” model and as a by-product estimates the likely future performance of this model while documenting to the researcher that there is usually great instability in the form of this model. It is imperative that the analyst repeat *all* feature selection or model selection steps afresh for each resample and displays the variation of features / models selected across the resamples. For the bootstrap this usually involves 300 or more resamples, and for cross-validation 50 or more repeats of 10-fold cross-validation. The “final” model should be describe as an example of an entire array of models, and the likely future performance of this example model is estimated from the resampling procedure just described. Resampling alerts the analyst alert to arbitrariness and reduces the tendency to cherry pick good validations when split-sample or external validation is used.

Thus in the case just described where a statistical model is not fully pre-specified, pretending to “freeze” the “final” result and validating it on a holdout sample is problematic. The resulting validation, far from being a validation of “the” model is just an example validation of an example model. On the other hand, rigorous validation using resampling validates the process used to derive the final predictive instrument, while still providing a good estimate of the likely future predictive accuracy of that instrument.

10.11.1

Summary: Choosing Internal vs. External Validation

Recall that strong internal validation uses the bootstrap in a way that repeats all modeling steps (feature/variable selection, transformation selection, parameter estimation, etc.) that utilized the outcome variable Y .

Use strong internal validation if

R

- the model is not pre-specified. If any feature selection utilizing Y was done, the set of features selected will be unstable, so an external validation would just validate an “example model.” On the other hand, a strong internal validation validates the model development *process* and fully documents the volatility of feature selection.
- the data come from multiple locations or times and you want to understand time trends in Y , or geographic differences
- the size of the potential external validation sample and the size of the training sample are not both very large

Use external validation if

S

- measurement platforms vary (e.g., genotyping or blood analysis equipment) and you want to validate the generalizability of a model developed on your platform
- both the training data and the test datasets are very large (e.g., the training data contains more than 15 events per **candidate** predictor and the test data has at least 200-400 events)
- the dataset to be used for the external validation was not available when the model

was developed and internally validated

- you don't trust the model developers to honestly perform a strong internal validation

10.11.2

Other Resources

T

- [Prediction Research Manual](#) by Cecile Janssens and Forike Martens
- Steyerberg paper on the waste by data splitting⁹⁹.

Chapter 11

Multiple Groups

11.1

Examples

- Compare baseline characteristics (e.g. age, height, BMI) or study response variable among subjects enrolled in one of three (or more) nonrandomized clinical trial arms
- Determine if pulmonary function, as measured by the forced expiratory volume in one second, differs in non-smokers, passive smokers, light smokers, and heavy smokers
- Evaluate differences in artery dilation among wild type, knockout, and knock-out+treated mice
 - Could add second factor: Normoxic (normal oxygen) or hypoxic (insufficient oxygen) environmental conditions (a *two-way* ANOVA)
- In general, studies with a continuous outcome and categorical predictors

11.2

The k -Sample Problem

- When $k = 2$ we compare two means or medians, etc.
- When $k > 2$ we could do all possible pairwise 2-sample tests but this can be misleading and may ↑ type I error
- Advantageous to get a single statistic testing H_0 : all groups have the same distribution (or at least the same central tendency)

11.3**Parametric ANOVA****11.3.1****Model**

- Notation
 - k groups (samples) each from a normal distribution
 - Population means $\mu_1, \mu_2, \dots, \mu_k$
 - n_i observations from the i th group
 - y_{ij} is the j th observation from i th group
- Model specification
 - $y_{ij} = \mu + \alpha_i + e_{ij}$
 - μ is a constant
 - α_i is a constant specific to group i
 - e_{ij} is the error term, which is assumed to follow a Normal distribution with mean 0 and variance σ^2
 - This model is overparameterized; that is, it is not possible to estimate μ and each α_i (a total of $k + 1$) terms using only k means
- Restriction 1: $\sum \alpha_i = 0$
 - μ is the mean of all groups taken together, the grand or overall mean
 - each α_i represents the deviation of the mean of the i th group from the overall mean
 - ϵ_{ij} is the deviation of individual data points from $\mu + \alpha_i$

- Restriction 2: $\alpha_1 = 0$
 - μ is the mean of group 1
 - each α_i represents the deviation of the mean of the i th group from the group 1 mean
 - ϵ_{ij} is the deviation of individual data points from $\mu + \alpha_i$
- Other restrictions possible, and will vary by software package

11.3.2

Hypothesis test

- Hypothesis test
 - $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
 - $H_1 : \text{at least two of the population means differ}$
- Not placing more importance on any particular pair or combination although large samples get more weight in the analysis
- Assume that each of the k populations has the same σ
- If $k = 2$ ANOVA yields identical P -value as 2-tailed 2-sample t -test
- ANOVA uses an F statistic and is always 2-tailed
- F ratio is proportional to the sum of squared differences between each sample mean and the grand mean over samples, divided by the sum of squared differences between all raw values and the mean of the sample from which the raw value came
- This is the SSB/SSW (sum of squares between / sum of squares within)
- SSB is identical to regression sum of squares
 SSW is identical to sum of squared errors in regression
- $F = MSB/MSW$ where

- MSB = mean square between = $SSB/(k - 1)$, $k - 1$ = “between group d.f.”
- MSW = mean square within = $SSW/(n - k)$, $n - k$ = “within group d.f.”
- Evidence for different $\mu_s \uparrow$ when differences in sample means (ignoring direction) are large in comparison to between-patient variation

11.3.3

Motivating Example

- Example from *Biostatistics: A methodology for the Health Sciences* by Fisher and Van Belle
- Research question (Zelazo et al., 1972, *Science*)
 - Outcome: Age at which child first walks (months)
 - Experiment involved the reinforcement of the walking and placing reflexes in newborns
 - Newborn children randomly assigned to one of four treatment groups
 - * Active exercise: Walking and placing stimulation 4 times a day for 8 weeks
 - * Passive exercise: An equal amount of gross motor stimulation
 - * No exercise: Tested along with first two groups at weekly intervals
 - * Control group: Infants only observed at 8 weeks (control for effect of repeated examinations)
- Distribution of ages (months) at which infants first walked alone. Data from Zelazo et al., 1972

	Active Group	Passive Group	No Exercise	8-week Control
	9.00	11.00	11.50	13.25
	9.50	10.00	12.00	11.50
	9.75	10.00	9.00	12.00
	10.00	11.75	11.50	13.50
	13.00	10.50	13.25	11.50
	9.50	15.00	13.00	12.35
Mean	10.125	11.375	11.708	12.350
Variance	2.0938	3.5938	2.3104	0.7400
Sum of Y_i	60.75	68.25	70.25	74.10

```
w <- rbind(
  data.frame(trt='Active', months=c(9,9.5,9.75,10,13,9.5)),
  data.frame(trt='Passive', months=c(11,10,10,11.75,10.5,15)),
  data.frame(trt='No Exercise', months=c(11.5,12,9,11.5,13.25,13)),
  data.frame(trt='8-Week Control', months=c(13.25,11.5,12,13.5,11.5,12.35))
)
aggregate(months ~ trt, w, function(x) c(Mean=mean(x), Variance=var(x)))
```

	trt	months.Mean	months.Variance
1	Active	10.125000	2.093750
2	Passive	11.375000	3.593750
3	No Exercise	11.708333	2.310417
4	8-Week Control	12.350000	0.740000

```
require(ggplot2)
require(data.table)
```

```
w <- data.table(w)
stats <- w[, j=list(months = mean(months), var=var(months)), by = trt]

ggplot(w, aes(x=trt, y=months)) +      # Fig. 11.1
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
  geom_errorbar(aes(ymin=..y.., ymax=..y..), width=.7, size=1.3,
                 data=stats) +
  xlab('') + ylab('Months Until First Walking') + coord_flip()
```

- Note that there are equal samples size in each group ($n_i = 6$ for each i) in the example. In general, this is not necessary for ANOVA, but it simplifies the calculations.
- Thought process for ANOVA
 - Assume that age at first walk is Normally distributed with some variance σ^2
 - The variance, σ^2 , is unknown. There are two ways of estimating σ^2

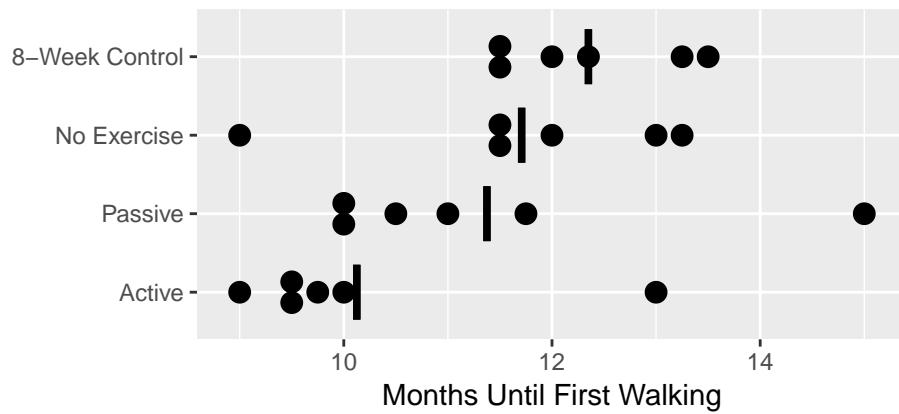


Figure 11.1: Age in months when infants first began walking by treatment group with mean lines

- Let the means in the four groups be μ_1, μ_2, μ_3 , and μ_4
- Method 1
 - * Assuming the variance are equal, calculated a pooled (or average) estimate of the variance using the four groups
 - * $s_p^2 = \frac{1}{4}(2.0938 + 3.5938 + 2.3104 + 0.7400) = 2.184$
- Method 2
 - * Assuming the four treatments do not differ ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$), the sample means follow a Normal distribution with variance $\sigma^2/6$.
 - * We can then estimated $\sigma^2/6$ by the variance of the sample means ($s_{\bar{y}}^2$)
 - * $s_{\bar{y}}^2 = \text{variance of } 10.125, 11.375, 11.708, 12.350$
 - * $s_{\bar{y}}^2 = 0.87349$, so $6s_{\bar{y}}^2 = 5.247$ is our second estimate of σ^2
 - s_p^2 is an estimate of the *within* group variability
 - $s_{\bar{y}}^2$ is an estimate of the *among* (or *between*) group variability
 - If H_0 is not true, method 2 will *overestimate* the variance
 - The hypothesis test is based on $F = 6s_{\bar{y}}^2/s_p^2$ and rejects H_0 if F is too large
- Degrees of Freedom
 - The F statistic has both a numerator and denominator degrees of freedom

- For the numerator, d.f. = $k - 1$
 - * There are k parameters $(\alpha_1, \alpha_2, \dots, \alpha_k)$
 - * And *one* restriction ($\sum \alpha_k = 0$, or $\alpha_1 = 0$, or another)
- For the denominator, d.f. = $N - k$
 - * There are N total observations
 - * And we estimate k sample means
- In the age at first walking example, there are 3 (numerator) and 20 (denominator) degrees of freedom

```
require(rms)
```

```
f <- ols(months ~ trt, data=w)
anova(f)
```

Analysis of Variance						
Factor	d.f.	Partial SS	MS	F	P	Response: months
trt	3	15.74031	5.246771	2.4	0.0979	
REGRESSION	3	15.74031	5.246771	2.4	0.0979	
ERROR	20	43.68958	2.184479			

11.3.4

Connection to Linear Regression

- Can do ANOVA using multiple regression, using an intercept and $k - 1$ “dummy” variables indicating group membership, so memorizing formulas specific to ANOVA is not needed
- Why is between group d.f.= $k - 1$?
 - can pick any one group as reference group, e.g., group 1
 - H_0 is identical to $H_0 : \mu_2 - \mu_1 = \mu_3 - \mu_1 = \dots = \mu_k - \mu_1 = 0$
 - if $k - 1$ differences in means are all zero, all means must be equal
 - since any unique $k - 1$ differences define our goal, there is $k - 1$ d.f. between groups for H_0

11.4

Why All These Distributions?

- Normal distribution is handy for approximating the distribution of z ratios (mean minus hypothesized value / standard error of mean) when n is large or σ is known
- If z is normal, z^2 has a χ_1^2 distribution
- If add k z^2 values the result has a χ_k^2 distribution; useful
 - in larger than 2×2 contingency tables
 - in testing goodness of fit of a histogram against a theoretical distribution
 - when testing more than one regression coefficient in regression models not having a σ to estimate
- t distribution: when σ is estimated from the data; exact P -values if data from normal population
Distribution indexed by d.f.: t_{df} ; useful for
 - testing one mean against a constant
 - comparing 2 means
 - testing one regression coefficient in multiple linear regression
- t_{df}^2 has an F distribution
- F statistic can test
 - > 1 regression coefficient
 - > 2 groups
 - whether ratio of 2 variances=1.0 (this includes MSB/MSW)
- To do this F needs two different d.f.
 - numerator d.f.: how many unique differences being tested (like χ_k^2)

- denominator d.f.
 - * total sample size minus the number of means or regression coefficients and intercepts estimated from the data
 - * is the denominator of the estimate of σ^2
 - * also called the error or residual d.f.
- $t_{df}^2 = F_{1,df}$
- ANOVA results in $F_{k-1,df}$; d.f.= $N - k$ where N = combined total sample size; cf. 2-sample t -test: d.f.= $n_1 + n_2 - 2$
- Example:

$$F = MSB/MSW = 58 \sim F_{4,1044}$$

The cumulative probability of getting an F statistic ≤ 58 with the above d.f. is 1.0000. We want $\text{Prob}(F \geq 58)$, so we get $P = 1 - 1 = 0$ to several digits of accuracy but report $P < 0.0001$.

```
pf(58, 4, 1044)
```

```
[1] 1
```

```
1 - pf(58, 4, 1044)
```

```
[1] 0
```

11.5

Software and Data Layout

- Every general-purpose statistical package does ANOVA
- Small datasets are often entered using Excel
- Statistical packages expect a grouping variable, e.g., a column of treatment names or numbers; a column of response values for all treatments combined is also present
- If you enter different groups' responses in different spreadsheets or different columns within a spreadsheet, it is harder to analyze the data with a stat package

11.6

Comparing Specific Groups

- F test is for finding any differences but it does not reveal which groups are different
- Often it suffices to quote F and P , then to provide sample means (and their confidence intervals)
- Can obtain CLs for any specific difference using previously discussed 2-sample t -test, but this can result in inconsistent results due solely to sampling variability in estimating the standard error of the difference in means using only the two groups to estimate the common σ
- If assume that there is a common σ , estimate it using all the data to get a pooled s^2
- $1 - \alpha$ CL for $\mu_i - \mu_j$ is then

$$\bar{y}_i - \bar{y}_j \pm t_{n-k,1-\alpha/2} \times s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}},$$

where n is the grand total sample size and there are respectively n_i and n_j observations in samples i and j

- Can test a specific $H_0 : \mu_i = \mu_j$ using similar calculations; Note that the d.f. for t comes from the grand sample size n , which \uparrow power and \downarrow width of CLs slightly
- Many people use more stringent α for individual tests when testing more than one of them (Section 11.10)
 - This is not as necessary when the overall F -test is significant

11.7

Non-Parametric ANOVA: Kruskal-Wallis Test

- k -sample extension to the 2-sample Wilcoxon–Mann–Whitney rank-sum test
- Is very efficient when compared to parametric ANOVA even if data are from normal distributions
- Has same benefits as Wilcoxon (not harmed by outliers, etc.)
- Almost testing for equality of population medians
- In general, tests whether observations in one group tend to be larger than observations in another group (when consider randomly chosen pairs of subjects)
- Test statistic obtained by replacing all responses by their ranks across all subjects (ignoring group) and then doing an ANOVA on the ranks
- Compute F (many authors use a χ^2 approximation but F gives more accurate P -values)
- Look up against the F distribution with $k - 1$ and $n - k$ d.f.
- Very accurate P -values except with very small samples
- Example:
 F statistic from ranks in Table 12.16: $F_{3,20} = 7.0289$
 - Using the cumulative distribution calculator from the web page, the prob. of getting an F less impressive than this under H_0 is 0.9979
 P is $1 - 0.9979 = 0.0021$
 - Compare with Rosner's $\chi^2_3 = 11.804$ from which $P = 0.008$ by survstat or one minus the CDF
 - Evidence that not all of the 4 samples are from the same distribution
 - loosely speaking, evidence for differences in medians

- better: some rabbits have larger anti-inflammatory effects when placed on different treatments in general
- Comparison of Kruskal-Wallis and Parametric ANOVA for age at first walk example
 - A few extreme values in age at first walk may violate parametric F -test assumptions
 - Run rank ANOVA: Kruskal-Wallis test three different ways:
 - * Parametric ANOVA on the ranks of y
 - * Spearman's ρ^2 generalized to multiple columns of x
 - * An R function dedicated to Kruskal-Wallis

```
anova(ols(rank(months) ~ trt, data=w))
```

Analysis of Variance						Response: rank(months)	
Factor	d.f.	Partial SS	MS	F	P		
trt	3	359.3333	119.77778	3.07	0.0515		
REGRESSION	3	359.3333	119.77778	3.07	0.0515		
ERROR	20	781.1667	39.05833				

```
spearman2(months ~ trt, data=w)
```

Spearman rho^2 Response variable:months						
rho2	F	df1	df2	P	Adjusted rho2	n
trt	0.315	3.07	3	0.0515	0.212	24

```
kruskal.test(months ~ trt, data=w)
```

Kruskal-Wallis rank sum test							
data: months by trt							
Kruskal-Wallis chi-squared = 7.2465, df = 3, p-value = 0.06444							

Note that the classical Kruskal-Wallis test uses the χ^2 approximation while the other two used an F distribution, which is as or more accurate than using χ^2 .

11.8

Two-Way ANOVA

- Ideal for a factorial design or observational study with 2 categorical grouping variables
- Example: 3 treatments are given to subjects and the researcher thinks that females and males will have different responses in general
Six means: $\bar{Y}_{i,j}$, $i = \text{treatment}$, $j = \text{sex group}$
- Can test
 - whether there are treatment differences after accounting for sex effects
 - whether there are sex differences after accounting for treatment effects
 - whether the treatment effect is difference for females and males, if allow treatment \times sex interaction to be in the model
- Suppose there are 2 treatments (A, B) and the 4 means are $\bar{Y}_{Af}, \bar{Y}_{Bf}, \bar{Y}_{Am}, \bar{Y}_{Bm}$, where f, m index the sex groups
- The various effects are estimated by
 - treatment effect: $\frac{(\bar{Y}_{Af} - \bar{Y}_{Bf}) + (\bar{Y}_{Am} - \bar{Y}_{Bm})}{2}$
 - sex effect: $\frac{(\bar{Y}_{Af} - \bar{Y}_{Am}) + (\bar{Y}_{Bf} - \bar{Y}_{Bm})}{2}$
 - treatment \times sex interaction: $(\bar{Y}_{Af} - \bar{Y}_{Bf}) - (\bar{Y}_{Am} - \bar{Y}_{Bm}) = (\bar{Y}_{Af} - \bar{Y}_{Am}) - (\bar{Y}_{Bf} - \bar{Y}_{Bm})$
- Interactions are “double differences”
- Assessing whether treatment effect is same for m vs. f is the same as assessing whether the sex effect is the same for A vs. B
- **Note:** 2-way ANOVA is **not** appropriate when one of the categorical variables represents conditions applied to the same subjects, e.g. serially collected data

within patient with time being one of the variables;
2-way ANOVA assumes that all observations come from different subjects

11.9

Analysis of Covariance

- Generalizes two-way ANOVA
- Allows adjustment for continuous variables when comparing groups
- Can ↑ power and precision by reducing unexplained patient to patient variability (σ^2)
- When Y is also measured at baseline, adjusting for the baseline version of Y can result in a major reduction in variance
- Fewer assumptions if adjust for baseline version of Y using ANCOVA instead of analyzing ($Y - \text{baseline } Y$)
- Two-way ANOVA is a special case of ANCOVA where a categorical variable is the only adjustment variable (it is represented in the model by dummy variables)

See Chapter 13 for much more information about ANCOVA in RCTs.

11.10

Multiple Comparisons

- When hypotheses are prespecified and are few in number, don't need to correct P -values or α level in CLs for multiple comparisons
- Multiple comparison adjustments are needed with H_0 is effectively in the form
 - Is one of the 5 treatments effective when compared against control?
 - Of the 4 etiologies of disease in our patients, is the treatment effective in at least one of them?
 - Is the treatment effective in either diabetics, older patients, males, . . . , etc.?
 - Diabetics had the greatest treatment effect empirically; the usual P -value for testing for treatment differences in diabetics was 0.03
- Recall that the probability that at least one event out of events E_1, E_2, \dots, E_m occurs is the sum of the probabilities if the events are mutually exclusive
- In general, the probability of at least one event is \leq the sum of the probabilities of the individual events occurring
- Let the event be “reject H_0 when it is true”, i.e., making a type I error or false positive conclusion
- If test 5 hypotheses (e.g., 5 subgroup treatment effects) at the 0.05 level, the upper limit on the chance of finding one significant difference if there are no differences at all is $5 \times 0.05 = 0.25$
- This is called the *Bonferroni inequality*
- If we test each H_0 at the $\frac{\alpha}{5}$ level the chance of at least one false positive is no greater than α
- The chance of at least one false positive is the *experimentwise error probability* whereas the chance that a specific test is positive by chance alone is the *comparisonwise error probability*

- Instead of doing each test at the $\frac{\alpha}{m}$ level we can get a conservative adjusted P -value by multiplying an individual P -value by m^a
- Whenever $m \times P > 1.0$ report $P = 1.0$
- There are many specialized and slightly less conservative multiple comparison adjustment procedures. Some more complex procedures are actually more conservative than Bonferroni.
- Statisticians generally have a poor understanding about the need to not only adjust P -values but to adjust point estimates also, when many estimates are made and only the impressive ones (by P) are discussed. In that case point estimates are badly biased away from the null value. For example, the BARI study analyzed around 20 subgroups and only found a difference in survival between PTCA and CABG in diabetics. The hazard ratio for CABG:PTCA estimated from this group is far too extreme.

^aMake sure that m is the total number of hypotheses tested with the data, whether formally or informally.

Chapter 12

Statistical Inference Review

- Emphasize confidence limits, which can be computed from adjusted or unadjusted analyses, with or without taking into account multiple comparisons
- P -values can accompany CLs if formal hypothesis testing needed
- When possible construct P -values to be consistent with how CLs are computed

12.1

Types of Analyses

- Except for one-sample tests, all tests can be thought of as testing for an association between at least one variable with at least one other variable
- Testing for group differences is the same as testing for association between group and response
- Testing for association between two continuous variables can be done using correlation (especially for unadjusted analysis) or regression methods; in simple cases the two are equivalent
- Testing for association between group and outcome, when there are more than 2 groups which are not in some solid order^a means comparing a summary of the response between k groups, sometimes in a pairwise fashion

^aThe dose of a drug or the severity of pain are examples of ordered variables.

12.2

Covariate-Unadjusted Analyses

Appropriate when

- Only interested in assessing the relationship between a single X and the response, or
- Treatments are randomized and there are no strong prognostic factors that are measureable
- Study is observational and variables capturing confounding are unavailable (place strong caveats in the paper)

See 13.

12.2.1

Analyzing Paired Responses

Type of Response	Recommended Test	Most Frequent Test
binary	McNemar	McNemar
continuous	Wilcoxon signed-rank	paired t -test

12.2.2

Comparing Two Groups

Type of Response	Recommended Test	Most Frequent Test
binary	$2 \times 2 \chi^2$	χ^2 , Fisher's exact test
ordinal	Wilcoxon 2-sample	Wilcoxon 2-sample
continuous	Wilcoxon 2-sample	2-sample <i>t</i> -test
time to event ^a	Cox model ^b	log-rank ^c

^aThe response variable may be right-censored, which happens if the subject ceased being followed before having the event. The value of the response variable, for example, for a subject followed 2 years without having the event is 2+.

^bIf the treatment is expected to have more early effect with the effect lessening over time, an accelerated failure time model such as the lognormal model is recommended.

^cThe log-rank is a special case of the Cox model. The Cox model provides slightly more accurate *P*-values than the χ^2 statistic from the log-rank test.

12.2.3

Comparing > 2 Groups

Type of Response	Recommended Test	Most Frequent Test
binary	$r \times 2 \chi^2$	χ^2 , Fisher's exact test
ordinal	Kruskal-Wallis	Kruskal-Wallis
continuous	Kruskal-Wallis	ANOVA
time to event	Cox model	log-rank

12.2.4

Correlating Two Continuous Variables

Recommended: Spearman ρ

Most frequently seen: Pearson r

12.3

Covariate-Adjusted Analyses

- To adjust for imbalances in prognostic factors in an observational study or for strong patient heterogeneity in a randomized study
- Analysis of covariance is preferred over stratification, especially if continuous adjustment variables are present or there are many adjustment variables
 - Continuous response: multiple linear regression with appropriate transformation of Y
 - Binary response: binary logistic regression model
 - Ordinal response: proportional odds ordinal logistic regression model
 - Time to event response, possibly right-censored:
 - * chronic disease: Cox proportional hazards model
 - * acute disease: accelerated failure time model

Chapter 13

Analysis of Covariance in Randomized Studies

Hierarchy of Causal Inference for Treatment Efficacy

Let P_i denote patient i and the treatments be denoted by A and B . Thus P_2^B represents patient 2 on treatment B . \bar{P}_1 represents the average outcome over a sample of patients from which patient 1 was selected.

Design	Patients Compared
6-period crossover	P_1^A vs P_1^B (directly measure HTE)
2-period crossover	P_1^A vs P_1^B
RCT in identical twins	P_1^A vs P_1^B
group RCT	\bar{P}_1^A vs \bar{P}_2^B , $P_1 = P_2$ on avg
Observational, good artificial control	\bar{P}_1^A vs \bar{P}_2^B , $P_1 = P_2$ hopefully on avg
Observational, poor artificial control	\bar{P}_1^A vs \bar{P}_2^B , $P_1 \neq P_2$ on avg
Real-world physician practice	P_1^A vs P_2^B

13.1

Covariate Adjustment in Linear Models

If you fail to adjust for pre-specified covariates, the statistical model's residuals get larger. Not good for power, but incorporates the uncertainties needed for any possible random baseline imbalances.

F Harrell 2019



ancova-linear

A

- Model: $E(Y|X) = X\beta + \epsilon$
- Continuous response variable Y , normal residuals
- Statistical testing for baseline differences is scientifically incorrect (Altman & Doré 1990, Begg 1990, Senn 1994, Austin *et al.* 2010); as Bland and Altman stated⁹, statistical tests draw inferences about *populations*, and the population model here would involve a repeat of the randomization to the whole population hence balance would be perfect. Therefore the null hypothesis of no difference in baseline distributions between treatment groups is automatically true.
- If we are worried about baseline imbalance we need to search patient records for counter-balancing factors
- → imbalance is not the reason to adjust for covariates
- Adjust to gain efficiency by subtracting explained variation
- Relative efficiency of unadjusted treatment comparison is $1 - \rho^2$
- Unadjusted analyses yields unbiased treatment effect estimate

See datamethods.org/t/should-we-ignore-covariate-imbalance-and-stop-presenting-a-stratified-table-one-for-randomized-trials for a detailed discussion, including reasons not to even stratify by treatment in “Table 1.”

13.2

Covariate Adjustment in Nonlinear Models



13.2.1

Hidden Assumptions in 2×2 Tables

ancova-nonlin
B

- Traditional presentation of 2-treatment clinical trial with a binary response: 2×2 table
- Parameters: P_1, P_2 for treatments 1, 2
- Test of goodness of fit: H_0 : all patients in one treatment group have same probability of positive response (P_j constant)
- $\rightarrow H_0$: no risk factors exist
- Need to account for patient heterogeneity
- ⁶³ has a method for estimating the bias in unadjusted the log odds ratio and also has excellent background information

13.2.2

Models for Binary Response

C

- Model for probability of event must be nonlinear in predictors unless risk range is tiny
- Useful summary of relative treatment effect is the odds ratio (OR)
- Use of binary logistic model for covariate adjustment will result in an **increase** in the S.E. of the treatment effect (log odds ratio) (Robinson & Jewell,⁸⁴)
- But even with perfect balance, adjusted OR \neq unadjusted OR
- Adjusted OR will be greater than unadjusted OR

Example from GUSTO-I

 D

- Steyerberg, Bossuyt, Lee¹⁰⁰
- Endpoint: 30-day mortality (0.07 overall)
- 10,348 patients given accelerated *t*-PA
- 20,162 patients given streptokinase (SK)
- Means and Percentages

Characteristics of 30,000 GUSTO Patients

Baseline Characteristic	<i>t</i> -PA	SK
Age	61.0	60.9
Female	25.3	25.3
Weight	79.6	79.4
Height	171.1	171.0
Hypertension	38.2	38.1
Diabetes	14.5	15.1
Never smoked	29.8	29.6
High cholesterol	34.6	34.3
Previous MI	16.9	16.5
Hypotension	8.0	8.3
Tachycardia	32.5	32.7
Anterior MI	38.9	38.9
Killip class I	85.0	85.4
ST elevation	37.3	37.8

Unadjusted / Adj. Logistic Estimates

E

- With and without adjusting for 17 baseline characteristics

Unadjusted and Adjusted GUSTO Analyses			
Type of Analysis	Log OR	S.E.	χ^2
Unadjusted	-0.159	0.049	10.8
Adjusted	-0.198	0.053	14.0

- Percent reduction in odds of death: 15% vs. 18%

- -0.159 (15%) is a biased estimate
- Increase in S.E. more than offset by increase in treatment effect
- Adjusted comparison based on 19% fewer patients would have given same power as unadjusted test

```
load('gustomin.rda')
with(gustomin,
      plot(density(p.sk), xlim=c(0, .4), xlab='Baseline Expected Risk',
            ylab='Probability Density', main='')) # Fig. 13.1
```

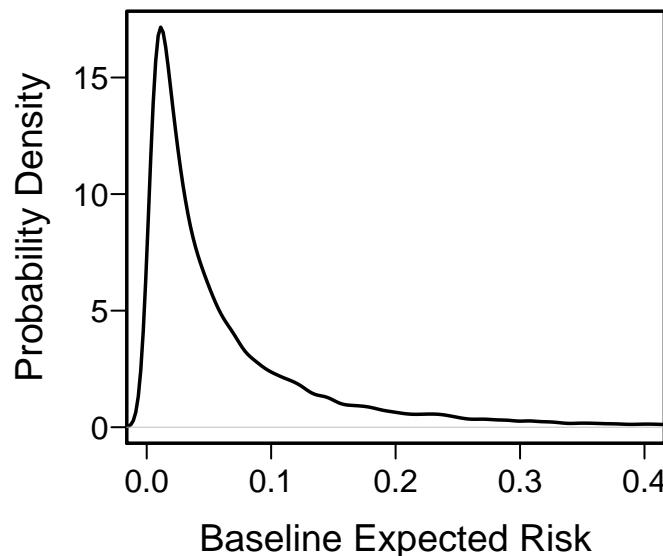


Figure 13.1: Distribution of baseline risk in GUSTO-I. Kernel density estimate of risk distribution for SK treatment. Average risk is 0.07. See also⁵⁰.

- Robinson & Jewell: “It is always more efficient to adjust for predictive covariates F when logistic models are used, and thus in this regard the behavior of logistic regression is the same as that of classic linear regression.”

Simple Logistic Example – Gail 1986

Male		
	Treatment A	Treatment B
$Y = 1$	500	100
$Y = 0$	500	900
	1000	1000
Odds Ratio: 9		

Female		
	Treatment A	Treatment B
$Y = 1$	900	500
$Y = 0$	100	500
		1000
		1000
Odds Ratio: 9		

Pooled		
	Treatment A	Treatment B
$Y = 1$	1400	600
$Y = 0$	600	1400
		2000
		2000
Odds Ratio: 5.44		

From seeing this example one can argue that odds ratios, like hazard ratios, were never really designed to be computed on a set of subjects having heterogeneity in their expected outcomes.

13.2.3

Nonlinear Models, General



- Gail, Wieand, Piantadosi³⁰ showed that for unadjusted treatment estimates to be unbiased, regression must be linear or exponential
- Gail³¹ showed that for logistic, Cox, and paired survival models unadjusted treatment effects are asymptotically biased low in absolute value
- Gail also studied normal, exponential, additive risk, Poisson

Part of the problem with Poisson, proportional hazard and logistic regression approaches is that they use a single parameter, the linear predictor, with no equivalent of the variance parameter in the Normal case. This means that lack of fit impacts on the estimate of the predictor.

Senn [91], p. 3747

13.3

Cox / Log-Rank Test for Time to Event

 ancova-efficiency
H

- Lagakos & Schoenfeld 1984 showed that type I error is preserved if don't adjust
- If hazards are proportional conditional on covariates, they are not proportional if omit covariates
- Morgan 1986 derived asymptotic relative efficiencies (ARE) of unadjusted log-rank test if a binary covariate is omitted
- If prevalence of covariate X is 0.5:

	$X = 1 : X = 0$	Hazard Ratio	ARE
		1.0	1.00
Efficiency of Unadjusted Log-Rank Test		1.5	0.95
		2.0	0.88
		3.0	0.72

I

- Ford, Norrie, Ahmadi²⁹: Treatment effect does not have the same interpretation under unadjusted and adjusted models
- No reason for the two hazard ratios to have the same value
- Akazawa, Nakamura, Palesch²: Power of unadjusted and stratified log-rank test

	Number of Strata	Range of Log Hazards	Power	
			Unadj.	Adjusted
	1	0	.78	-
Power With and Without Adjustment	2	0–0.5	.77	.78
		0–1	.67	.78
		0–2	.36	.77
	4	0–3	.35	.77
	8	0–3.5	.33	.77

13.3.1

Sample Size Calculation Issues

J

- Schoenfeld⁸⁹ implies that covariate adjustment can only ↑ sample size in randomized trials
- Need to recognize ill-definition of unadjusted hazard ratios

13.4

Why are Adjusted Estimates Right?



ancova-meaning
K

- Hauck, Anderson, Marcus⁴³, who have an excellent review of covariate adjustment in nonlinear models, state:

"For use in a clinician–patient context, there is only a single person, that patient, of interest. The subject-specific measure then best reflects the risks or benefits for that patient. Gail has noted this previously [ENAR Presidential Invited Address, April 1990], arguing that one goal of a clinical trial ought to be to predict the direction and size of a treatment benefit for a patient with specific covariate values. In contrast, population–averaged estimates of treatment effect compare outcomes in groups of patients. The groups being compared are determined by whatever covariates are included in the model. The treatment effect is then a comparison of average outcomes, where the averaging is over all omitted covariates."

13.5

How Many Covariables to Use?

- Try to adjust for the bulk of the variation in outcome^{43,105}
- Neuhaus⁷³: “to improve the efficiency of estimated covariate effects of interest, analysts of randomized clinical trial data should adjust for covariates that are strongly associated with the outcome”
- Raab *et al.*⁸³ have more guidance for choosing covariables and provide a formula for linear model that shows how the value of adding a covariate depends on the sample size

ancova-plan
L

13.6

Differential and Absolute Treatment Effects



13.6.1

Modeling Differential Treatment Effect

Differential treatment effect is often called *heterogeneity of treatment effect* or HTE. As opposed to the natural expansion of absolute treatment effect with underlying subject risk, differential treatment effect is usually based on analyses of relative effects, especially when the outcome is binary.

blog

ancova-hte

The most common approach to analyzing differential treatment effect involves searching for such effects rather than estimating the differential effect. This is, tragically, most often done through subgroup analysis.

Problems With Subgroup Analysis

M

- Subgroup analysis is widely practiced and widely derided in RCTs (as well as in observational studies)
- Construction of a subgroup from underlying factors that are continuous in nature (e.g., “older” = age ≥ 65) assumes that the treatment effect is like falling off a cliff, i.e., all-or-nothing. Discontinuous treatment effects, like discontinuous main effects, have not been found and validated but have always been shown to have been an artificial construct driven by opinion and not data.
- Given a subgroup a simple label such as “class IV heart failure” may seem to be meaningful but subgrouping carries along other subject characteristics that are correlated with the subgrouping variable. So the subgroup’s treatment effect estimate has a more complex interpretation than what is thought.
- Researchers who don’t understand that “absence of evidence is not evidence for absence” interpret a significant effect in one subgroup but not in another as the treatment being effective in the first but ineffective in the second. The second

P-value may be large because of increased variability in the second subgroup or because of a smaller effective sample size in that group.

- Testing the treatment effect in a subgroup does not carry along with it the covariate adjustment needed and assumes there is no residual HTE *within* the subgroup.
- When treatment interacts with factors not used in forming the subgroup, or when the subgrouping variable interacts with an ignored patient characteristic, the subgroup treatment effect may be grossly misleading. As an example, in GUSTO-I there was a strong interaction between Killip class and age. Unadjusted analysis of treatment effect within older subjects was partially an analysis of Killip class. And the treatment effect was not “significant” in older patients, leading many readers to conclude *t*-PA should not be given to them. In fact there was no evidence that age interacted with treatment, and the absolute risk reduction due to *t*-PA *increased* with age.

Specifying Interactions

N
O

- Assessing differential treatment effect best done with formal interaction tests rather than subgroup analysis
- Pre-specify sensible effect modifiers
 - interactions between treatment and extent of disease
 - “learned” interventions: interaction between treatment and duration of use by physician
- Interactions with center are not necessarily sensible
- Need to use penalized estimation (e.g., interaction effects as random effects) to get sufficient precision of differential treatment effects, if # interaction d.f. > 4 for example^{88,115}

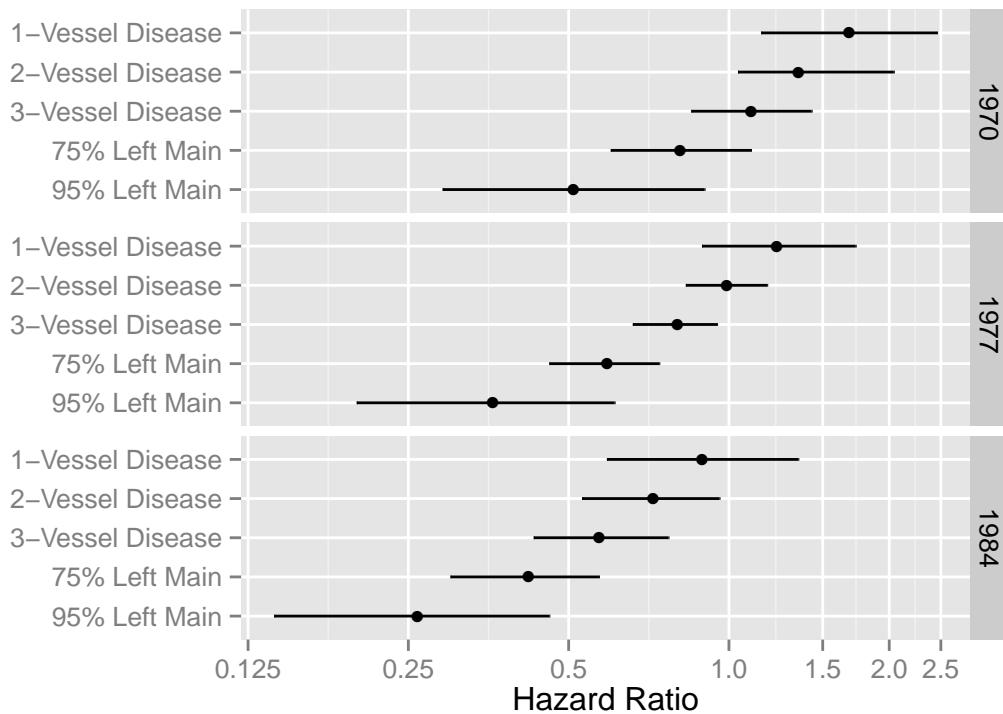


Figure 13.2: A display of an interaction between treatment, extent of disease, and calendar year of start of treatment¹³

Strategy for Analyzing Differential Treatment Effect



The broad strategy that is recommended is based on the following:

P

- Anything resembling subgroup analysis should be avoided
- Anything that assumes that the treatment effect has a discontinuity in a continuous variable should be avoided. Differential effects should be smooth dose-response effects.
- Anything that results in a finding that has a simpler alternate explanation should be avoided
 - Specify models so that an apparent interaction is not just a stand-in for an omitted main effect

Particulars of the strategy are:

Q

- Formulate a subject-matter driven covariate model, including all covariates understood to have strong effects on the outcome. Ensure that these covariates are adjusted for in every HTE analysis context.
- Main effects need to be flexibly modeled (e.g., using regression splines) so as to not assume linearity. False linearity assumptions can inflate apparent interaction effects because the interaction may be co-linear with omitted nonlinear main effects.
- If the sample size allows, also model interaction effects as smooth nonlinear functions. As with main effects, it is not uncommon for interaction effects to be nonlinear, e.g., the effect of treatment is small for age < 70 then starts to expand rapidly after age 70^a. As a compromise, force interactions to be linear even if main effects are not, if the effective sample size does not permit estimating parameters for nonlinear differential effects.
- Consider effect modifiers that are somewhat uncorrelated with each other. Add the main effects corresponding to each potential effect modifier into the model described in the previous point. For example if the primary analysis uses a model containing age, sex, and severity of disease and one is interested in assessing HTE by race and by geographical region, add *both* race and region to the main effects that are adjusted for in *all* later models. This will result in proper covariate adjustment and will handle the case where an apparent interaction is partially explained by an omitted main effect that is co-linear with one of the interacting factors.
- Carry out a joint (chunk) likelihood ratio or F -test for all potential interaction effects combined. This test has a perfect multiplicity adjustment and will not be “brought down” by the potential interactions being co-linear with each other. The P -value from this joint test with multiple degrees of freedom will place the tests described in the next step in context. But note that with many d.f. the test may lack power. R
- Consider each potential interaction one-at-a-time by adding that interaction term to the *comprehensive* main effects model. In the example given above, main effects simultaneously adjusted for would include treatment, age, sex, severity of disease, race, and region. Then treatment \times race is added to the model and tested. Then the race interaction term (but not its main effect) is removed from the model and is replaced by region interactions.

^aThis does not imply that age should be modeled as a categorical variable to estimate the interaction; that would result in unexplained HTE in the age ≥ 70 interval.

Concerning the last step, we must really recognize that there are two purposes for the analysis of differential treatment effect:

S

1. Understanding which subject characteristics are associated with alterations in efficacy
2. Predicting the efficacy for an individual subject

For the latter purpose, one might include all potential treatment interactions in the statistical model, whether interactions for multiple baseline variables are co-linear or not. Such co-linearities do not hurt predictions or confidence intervals for predictions. For the former purpose, interactions can make one effect modifier compete with another, and neither modifier will be significant when adjusted for the other. As an example, suppose that the model had main effects for treatment, age, sex, and systolic and diastolic blood pressure (SBP, DBP) and we entertained effect modification due to one or both of SBP, DBP. Because of the correlation between SBP and DBP, the main effects for these variables are co-linear and their statistical significance is weakened because they estimate, for example, the effect of increasing SBP holding DBP constant (which is difficult to do). Likewise, treatment \times SBP and treatment \times DBP interactions are co-linear with each other, making statistical tests for interaction have low power. We may find that neither SBP nor DBP interaction is “significant” after adjusting for the other, but that the combined chunk test for SBP or DBP interaction is highly significant. Someone who does not perform the chunk test may falsely conclude that SBP does not modify the treatment effect. This problem is more acute when more than two factors are allowed to interact with treatment.

Note that including interaction terms in the model makes all treatment effect estimates conditional on specific covariate values, effectively lowering the sample size for each treatment effect estimate. When there is no HTE, the overall treatment main effect without having interaction terms in the model is by far the highest precision estimate. There is a bias-variance tradeoff when considering HTE. Adding interaction terms lowers bias but greatly increases variance.

Another way to state the HTE estimation strategy is as follows.

T

1. Pick a model for which it is mathematically possible that there be no interactions (no restrictions on model parameters)
2. Develop a model with ample flexibly-modeled main effects and clinically pre-specified interactions. Use a Bayesian skeptical prior or penalized maximum likelihood esti-

mate to shrink the interaction terms if the effective sample size does not support the pre-specified number of interaction parameters. Be sure to model interaction effects as continuous if they come from variables of an underlying continuous nature.

3. Display and explain relative effects (e.g., odds ratios) as a function of interacting factors and link that to what is known about the mechanism of treatment effect.
4. Put all this in a context (like Figure 13.4) that shows absolute treatment benefit (e.g., risk scale) as a function of background risk and of interacting factors.
5. State clearly that background risk comes from all non-treatment factors and is a risk-difference accelerator that would increase risk differences for any risk factor, not just treatment. Possibly provide a risk calculator to estimate background risk to plug into the x -axis.

Consider simulated two-treatment clinical trial data to illustrate the “always adjust for all main effects but consider interactions one at a time” approach to analyzing and displaying evidence for differential treatment effect. After that we will illustrate a simultaneous interaction analysis. Simulate time-to-event data from an exponential distribution, and fit Cox proportional hazards models to estimate the interaction between age and treatment and between sex and treatment. The true age effect is simulated as linear in the log hazard and the treatment effect on the log relative hazard scale is proportional to how far above 60 years is a patient’s age, with no treatment benefit before age 60. This is a non-simple interaction that could be exactly modeled with a linear spline function, but assume that the analyst does not know the true form of the interaction so she allows for a more general smooth form using a restricted cubic spline function. The data are simulated so that there is no sex interaction.

```
require(rms)

options(prType='latex')      # for cph print, anova
set.seed(1)
n <- 3000      # total of 3000 subjects
age <- rnorm(n, 60, 12)
label(age) <- 'Age'
sex <- factor(sample(c('Male', 'Female'), n, rep=TRUE))
treat <- factor(sample(c('A', 'B'), n, rep=TRUE))
cens <- 15 * runif(n)      # censoring time
h <- 0.02 * exp(0.04 * (age - 60) + 0.4 * (sex == 'Female') -
                0.04 * (treat == 'B') * pmax(age - 60, 0))
dt <- -log(runif(n)) / h
label(dt) <- 'Time Until Death or Censoring'
e <- ifelse(dt <= cens, 1, 0)
dt <- pmin(dt, cens)
units(dt) <- 'Year'
```

```
dd <- datadist(age, sex, treat); options(datadist='dd')
S <- Surv(dt, e)
f <- cph(S ~ sex + rcs(age, 4) * treat)
f
```

Cox Proportional Hazards Model

```
cph(formula = S ~ sex + rcs(age, 4) * treat)
```

		Model Tests		Discrimination Indexes	
Obs	3000	LR χ^2	130.78	R^2	0.047
Events	470	d.f.	8	D_{xy}	0.261
Center	2.3448	$Pr(> \chi^2)$	0.0000	g	0.566
		Score χ^2	156.15	g_r	1.761
		$Pr(> \chi^2)$	0.0000		

	$\hat{\beta}$	S.E.	Wald Z	$Pr(> Z)$
sex=Male	-0.3628	0.0936	-3.88	0.0001
age	0.0423	0.0283	1.50	0.1349
age'	0.0132	0.0655	0.20	0.8404
age"	-0.0456	0.2499	-0.18	0.8553
treat=B	1.6377	1.6902	0.97	0.3326
age × treat=B	-0.0335	0.0357	-0.94	0.3479
age' × treat=B	0.0643	0.0892	0.72	0.4709
age" × treat=B	-0.4048	0.3606	-1.12	0.2616

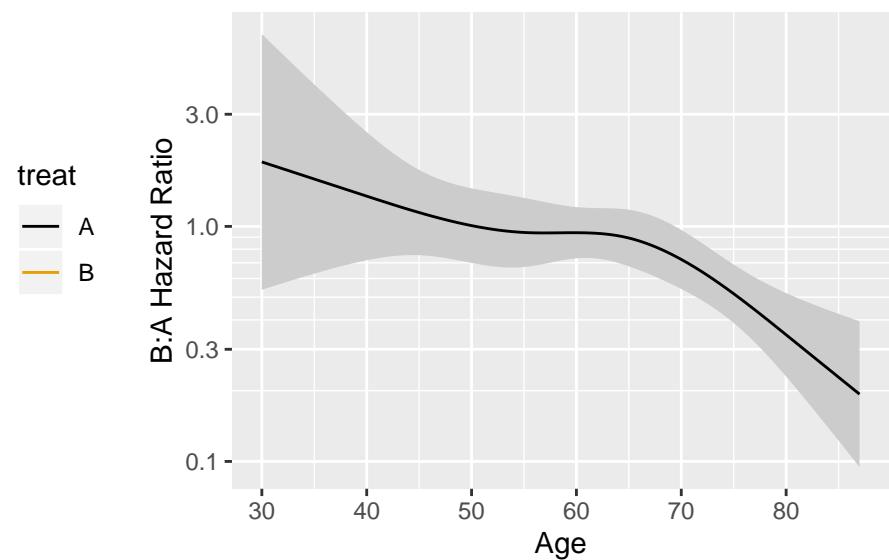
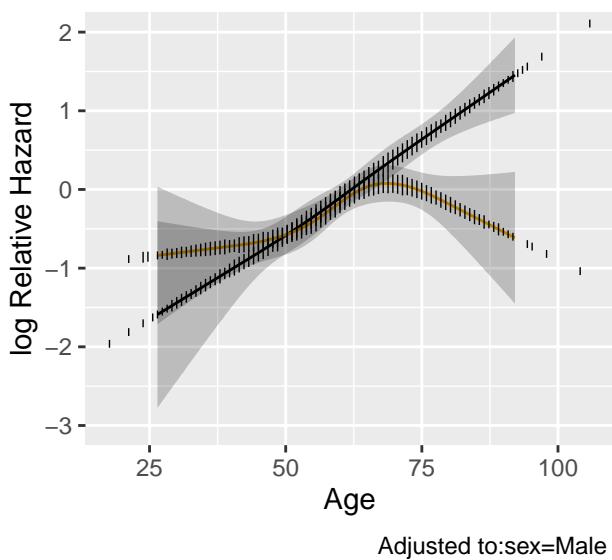
```
anova(f)
```

Wald Statistics for S

	χ^2	d.f.	P
sex	15.03	1	0.0001
age (Factor+Higher Order Factors)	102.95	6	<0.0001
<i>All Interactions</i>	19.92	3	0.0002
<i>Nonlinear (Factor+Higher Order Factors)</i>	6.04	4	0.1961
treat (Factor+Higher Order Factors)	27.80	4	<0.0001
<i>All Interactions</i>	19.92	3	0.0002
age \times treat (Factor+Higher Order Factors)	19.92	3	0.0002
<i>Nonlinear</i>	4.11	2	0.1284
<i>Nonlinear Interaction : f(A,B) vs. AB</i>	4.11	2	0.1284
TOTAL NONLINEAR	6.04	4	0.1961
TOTAL NONLINEAR + INTERACTION	20.80	5	0.0009
TOTAL	138.91	8	<0.0001

The model fitted above allows for a general age \times treatment interaction. Let's explore this interaction by plotting the age effect separately by treatment group, then plotting the treatment B:A hazard ratio as a function of age.

```
ggplot(Predict(f, age, treat), rdata=data.frame(age, treat))
ages <- seq(30, 87, length=200)
k <- contrast(f, list(treat='B', age=ages), list(treat='A', age=ages))
k <- as.data.frame(k[Cs(sex,age,Contrast,Lower,Upper)])
ggplot(k, aes(x=age, y=exp(Contrast))) +
  scale_y_log10(minor_breaks=seq(.2, .9, by=.1)) +
  geom_ribbon(aes(ymin=exp(Lower), ymax=exp(Upper)), fill='gray80') +
  geom_line() +
  ylab('B:A Hazard Ratio') + xlab('Age')
```



Re-fit the model allowing for a sex interaction (but not an age interaction) and display

the results.

```
g ← cph(S ~ sex * treat + rcs(age, 4))
g
```

Cox Proportional Hazards Model

```
cph(formula = S ~ sex * treat + rcs(age, 4))
```

	Model Tests		Discrimination Indexes	
Obs	3000	LR χ^2	108.40	R^2 0.039
Events	470	d.f.	6	D_{xy} 0.249
Center	1.2565	Pr(> χ^2)	0.0000	g 0.566
		Score χ^2	110.68	g_r 1.761
		Pr(> χ^2)	0.0000	

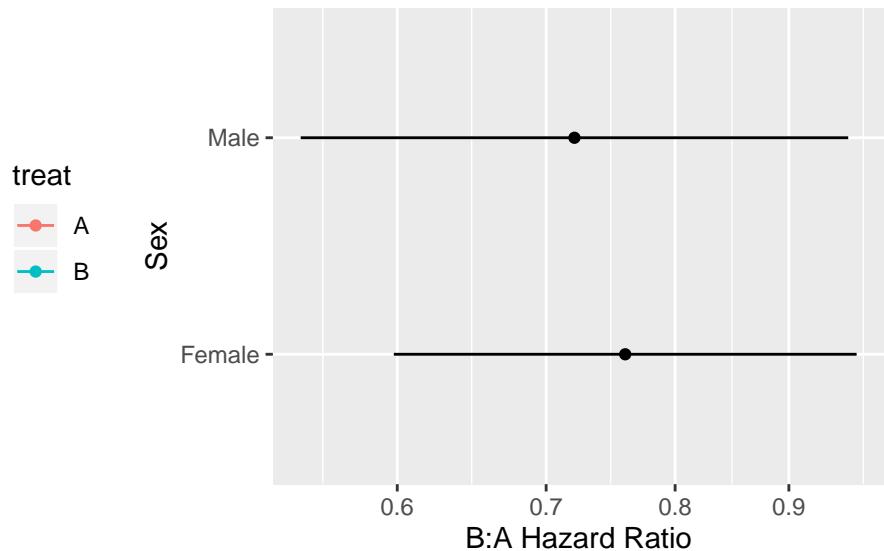
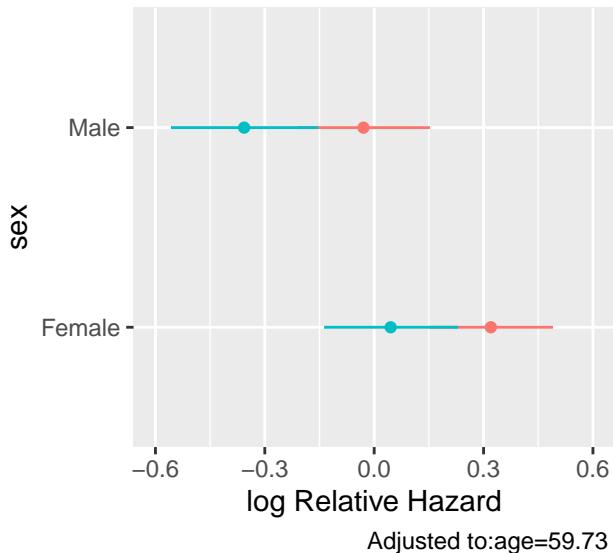
	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
sex=Male	-0.3494	0.1234	-2.83	0.0046
treat=B	-0.2748	0.1222	-2.25	0.0246
age	0.0228	0.0173	1.31	0.1890
age'	0.0452	0.0435	1.04	0.2981
age''	-0.2088	0.1739	-1.20	0.2301
sex=Male × treat=B	-0.0526	0.1894	-0.28	0.7813

```
anova(g)
```

Wald Statistics for S

	χ^2	d.f.	P
sex (Factor+Higher Order Factors)	15.87	2	0.0004
All Interactions	0.08	1	0.7813
treat (Factor+Higher Order Factors)	10.18	2	0.0062
All Interactions	0.08	1	0.7813
age	78.81	3	<0.0001
Nonlinear	1.85	2	0.3966
sex × treat (Factor+Higher Order Factors)	0.08	1	0.7813
TOTAL NONLINEAR + INTERACTION	1.92	3	0.5888
TOTAL	104.05	6	<0.0001

```
ggplot(Predict(g, sex, treat))
k <- contrast(g, list(treat='B', sex=levels(sex)), list(treat='A', sex=levels(
  sex)))
k <- as.data.frame(k[Cs(sex,age,Contrast,Lower,Upper)])
ggplot(k, aes(y=exp(Contrast), x=sex)) + geom_point() +
  scale_y_log10(breaks=c(.5, .6, .7, .8, .9, 1, 1.1)) +
  geom_linerange(aes(ymin=exp(Lower), ymax=exp(Upper))) +
  xlab('Sex') + ylab('B:A Hazard Ratio') + coord_flip()
```



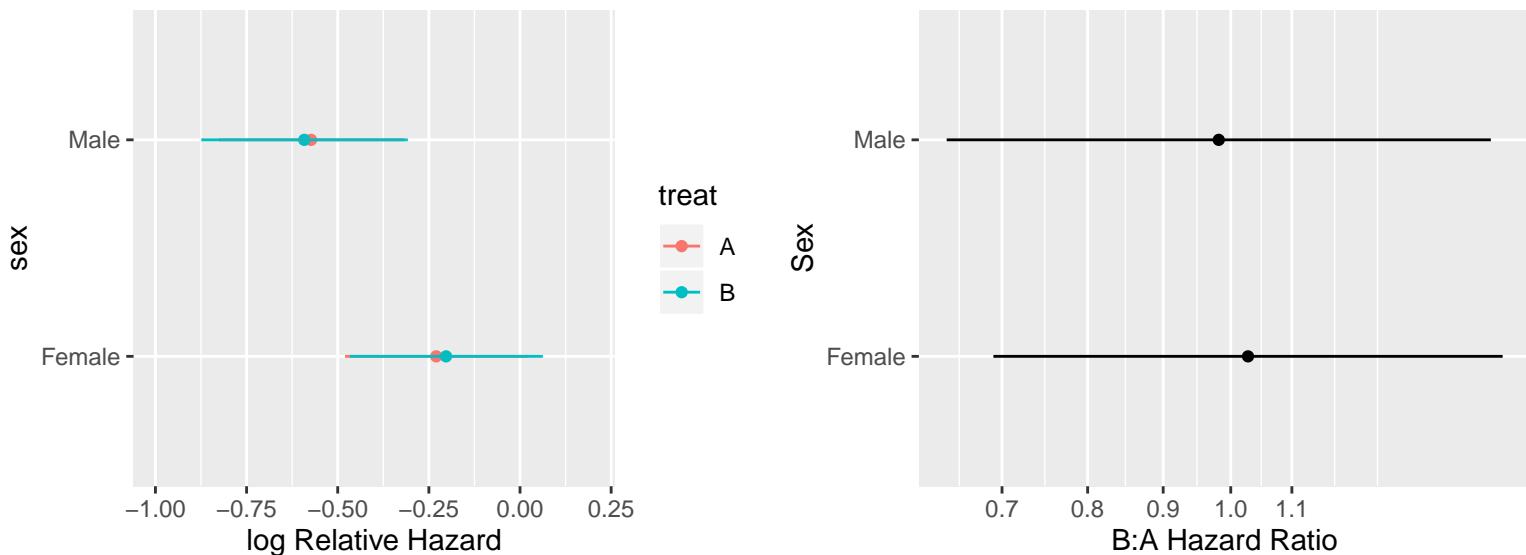
The above analysis signified a treatment effect for both sex groups (both confidence limits exclude a hazard ratio of 1.0), whereas this should only be the case when age exceeds 60. No sex \times treatment interaction was indicated. Re-do the previous analysis adjusting for an age \times treatment interaction while estimating the sex \times treatment interaction. For this purpose we must specify an age value when estimating the treatment effects by sex. Here we set age to 50 years. Were age and sex correlated, the joint analysis would have been harder to interpret.

```
h <- cph(S ~ treat * (rcs(age, 4) + sex))
anova(h)
```

Wald Statistics for S

	χ^2	d.f.	P
treat (Factor+Higher Order Factors)	27.85	5	<0.0001
All Interactions	19.98	4	0.0005
age (Factor+Higher Order Factors)	103.01	6	<0.0001
All Interactions	19.91	3	0.0002
Nonlinear (Factor+Higher Order Factors)	6.05	4	0.1956
sex (Factor+Higher Order Factors)	15.08	2	0.0005
All Interactions	0.06	1	0.8096
treat × age (Factor+Higher Order Factors)	19.91	3	0.0002
Nonlinear	4.10	2	0.1286
Nonlinear Interaction : $f(A,B)$ vs. AB	4.10	2	0.1286
treat × sex (Factor+Higher Order Factors)	0.06	1	0.8096
TOTAL NONLINEAR	6.05	4	0.1956
TOTAL INTERACTION	19.98	4	0.0005
TOTAL NONLINEAR + INTERACTION	20.86	6	0.0019
TOTAL	138.58	9	<0.0001

```
ggplot(Predict(h, sex, treat, age=50))
k <- contrast(h, list(treat='B', sex=levels(sex), age=50),
               list(treat='A', sex=levels(sex), age=50))
k <- as.data.frame(k[Cs(sex,age,Contrast,Lower,Upper)])
ggplot(k, aes(y=exp(Contrast), x=sex)) + geom_point() +
  scale_y_log10(breaks=c(.5, .6, .7, .8, .9, 1, 1.1)) +
  geom_linerange(aes(ymin=exp(Lower), ymax=exp(Upper))) +
  xlab('Sex') + ylab('B:A Hazard Ratio') + coord_flip()
```



This is a more accurate representation of the true underlying model, and is easy to interpret because (1) age and sex are uncorrelated in the simulation model used, and (2) only two interactions were considered.

Absolute vs. Relative Treatment Effects Revisited



ancova-absolute

Statistical models are typically chosen so as to maximize the likelihood of the model fitting the data and processes that generated them. Even though one may be interested in absolute effects, models must usually be based on relative effects so that quantities such as estimated risk are in the legal $[0, 1]$ range. It is not unreasonable to think of a relative effects model such as the logistic model as providing supporting evidence that an interacting effect *causes* a change in the treatment effect, whereas the necessary expansion of treatment effect for higher risk subjects, as detailed in the next section, is merely a mathematical necessity for how probabilities work. One could perhaps say that any factor that confers increased risk for a subject (up to the point of diminishing returns; see below) might *cause* an increase in the treatment effect, but this is a general phenomenon that is spread throughout all risk factors independently of how they may directly affect the treatment effect. Because of this, additive risk models are not recommended for modeling outcomes or estimating differential treatment effects on an absolute scale. This logic leads to the following recommended strategy:

U

1. Base all inference and predicted risk estimation on a model most likely to fit the data (e.g., logistic risk model; Cox model). Choose the model so that it is *possible* that there may be no interactions, i.e., a model that does not place a restriction on regression coefficients.
2. Pre-specify sensible possible interactions as described earlier
3. Use estimates from this model to estimate relative differential treatment effects, which should be smooth functions of continuous baseline variables
4. Use the same model to estimate absolute risk differences (as in the next section) as a function both of interacting factors and of baseline variables in the model that do not interact with treatment (but will still expand the absolute treatment effect)
5. Even better: since expansion of risk differences is a general function of overall risk and not of individual risk factors only display effects of individual risk factors that interact with treatment (interacting on the scale that would have allowed them not to interact—the relative scale such as log odds) and show the general risk difference expansion as in Figures 13.3 (think of the “risk factor” as treatment) and 13.6

To translate the results of clinical trials into practice may require a lot of work involving modelling and further background information. ‘Additive at the point of analysis but relevant at the point of application’ should be the motto.

Stephen Senn in <http://errorstatistics.com/2013/04/19/stephen-senn-when-relevance-is-irrelevant>

13.6.2

Estimating Absolute Treatment Effects



anova-absolute
v

- Absolute efficacy measures:
 - Risk difference (δ)
 - number needed to treat (reciprocal of risk difference)
 - Years of life saved
 - Quality-adjusted life years saved
- Binary response, no interactions with treatment, risk for control patient P :

$$\delta = P - \frac{P}{P + (1-P)/OR}$$
- δ is dominated by P

```
plot(0, 0, type="n", xlab="Risk for Subject Without Risk Factor",
     ylab="Increase in Risk",
     xlim=c(0,1), ylim=c(0,.6)) # Figure 13.3
i <- 0
or <- c(1.1,1.25,1.5,1.75,2,3,4,5,10)
for(h in or) {
  i <- i + 1
  p <- seq(.0001, .9999, length=200)
  logit <- log(p/(1 - p)) # same as qlogis(p)
  logit <- logit + log(h) # modify by odds ratio
  p2 <- 1/(1 + exp(-logit))# same as plogis(logit)
  d <- p2 - p
  lines(p, d, lty=i)
  maxd <- max(d)
  smax <- p[d==maxd]
  text(smax, maxd + .02, format(h), cex=.6)
}
```

w

If the outcome is such that $Y = 1$ implies a good outcome, Figure 13.3 would be useful for estimating the absolute risk increase for a “good” treatment by selecting the one curve according to the odds ratio the treatment achieved in a multivariable risk model. This assumes that the treatment does not interact with any patient characteristic(s).

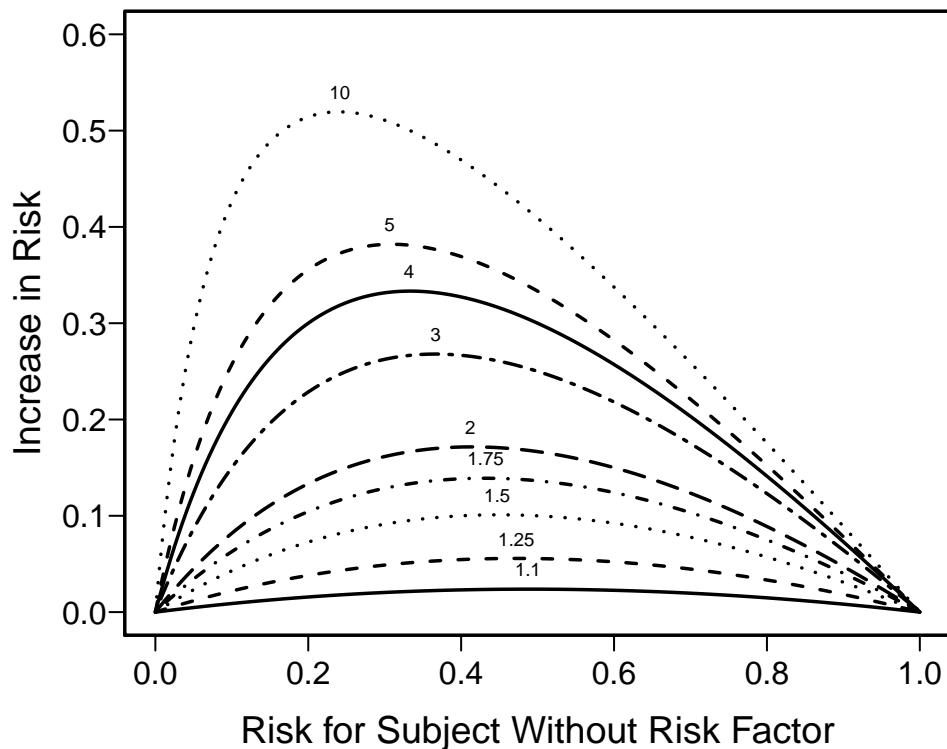


Figure 13.3: Absolute risk increase as a function of risk for control subject. Numbers on curves are treatment:control odds ratios.

van Klaveren et al.¹⁰⁷ described the importance of correctly modeling interactions when estimating absolute treatment benefit.

Now consider the case where $Y = 1$ is a bad outcome and $Y = 0$ is a good outcome, and there is differential relative treatment effect according to a truly binary patient characteristic $X = 0, 1$. Suppose that treatment represents a new agent and the control group is patients on standard therapy. Suppose that the new treatment multiplies the odds of a bad outcome by 0.8 when $X = 0$ and by 0.6 when $X = 1$, and that the background risk that $Y = 1$ for patients on standard therapy ranges from 0.01 to 0.99. The background risk could come from one or more continuous variables or mixtures of continuous and categorical patient characteristics. The “main effect” of X must also be specified. We assume that $X = 0$ goes into the background risk and $X = 1$ increases the odds that $Y = 1$ by a factor of 1.4 for patients on standard therapy. All of this specifies a full probability model that can be evaluated to show the absolute risk reduction by the new treatment as a function of background risk and X .

```

require(Hmisc)
d <- expand.grid(X=0:1, brisk=seq(0.01, 0.99, length=150))
d <- upData(d,
            risk.standard = plogis(qlogis(brisk) + log(1.4) * X),
            risk.new       = plogis(qlogis(brisk) + log(1.4) * X +
                                      log(0.8) * (X == 0) +
                                      log(0.6) * (X == 1)),
            .drop=TRUE)

```

```
risk.diff      = risk.standard - risk.new,
X = factor(X) )
```

Input object size:	19040 bytes;	2 variables	300 observations
Added variable	risk.standard		
Added variable	risk.new		
Added variable	risk.diff		
Modified variable	X		
New object size:	27152 bytes;	5 variables	300 observations

```
ggplot(d, aes(x=risk.standard, y=risk.diff, color=X)) +
  geom_line() +
  xlab('Risk Under Standard Treatment') +
  ylab('Absolute Risk Reduction With New Treatment') #
```

X

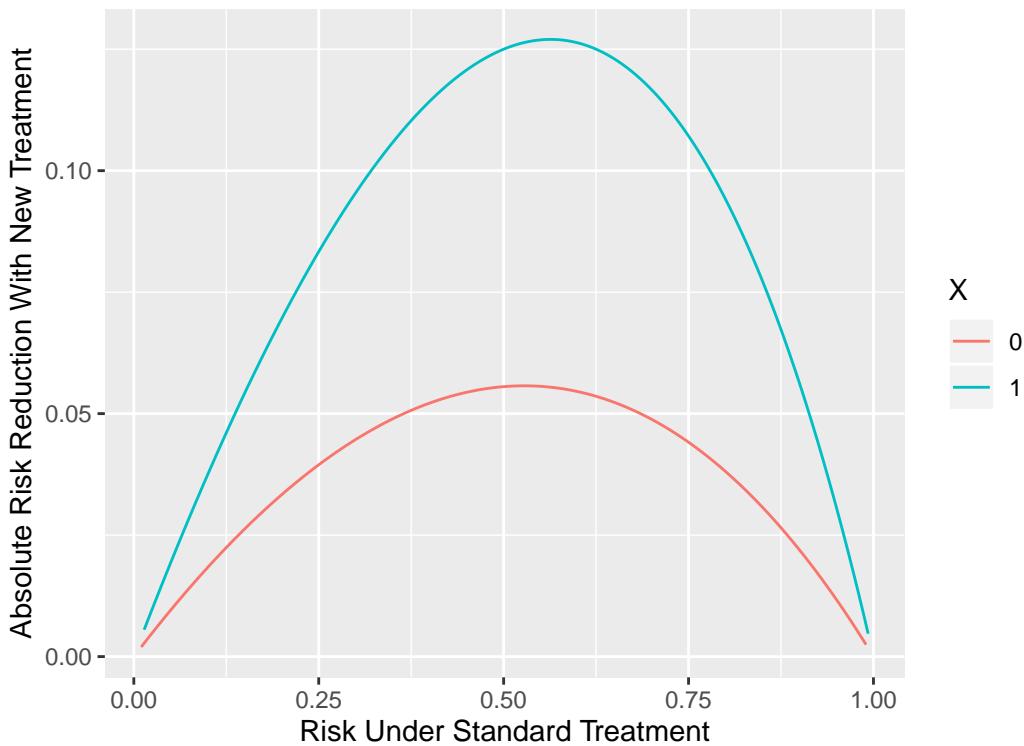


Figure 13.4: Absolute risk reduction by a new treatment as a function of background risk and an interacting factor

It is important to note that the magnification of absolute risk reduction by increasing background risk should not be labeled (or analyzed) by any one contributing risk factor. This is a generalized effect that comes solely from the restriction of probabilities to the $[0, 1]$ range.

Absolute Treatment Effects for GUSTO-I



- No evidence for interactions with treatment

- Misleading subgroup analysis showed that elderly patients not benefit from *t*-PA; result of strong age \times Killip class interaction
- Wide variation in absolute benefit of *t*-PA

```
delta <- with(gustomin, p.sk - p.tpa)
plot(density(delta), xlab='Mortality Difference',
      ylab='Probability Density', main='')    # Fig. 13.5
m <- mean(delta)
u <- par("usr")
arrows(m, u[3], m, 0, length=.1, lwd=2)
text(m, 2, 'Mean', srt=45, adj=0)
med <- median(delta)
arrows(med, u[3], med, 0, length=.1, lwd=2)
text(med, 2, 'Median', srt=45, adj=0)
```

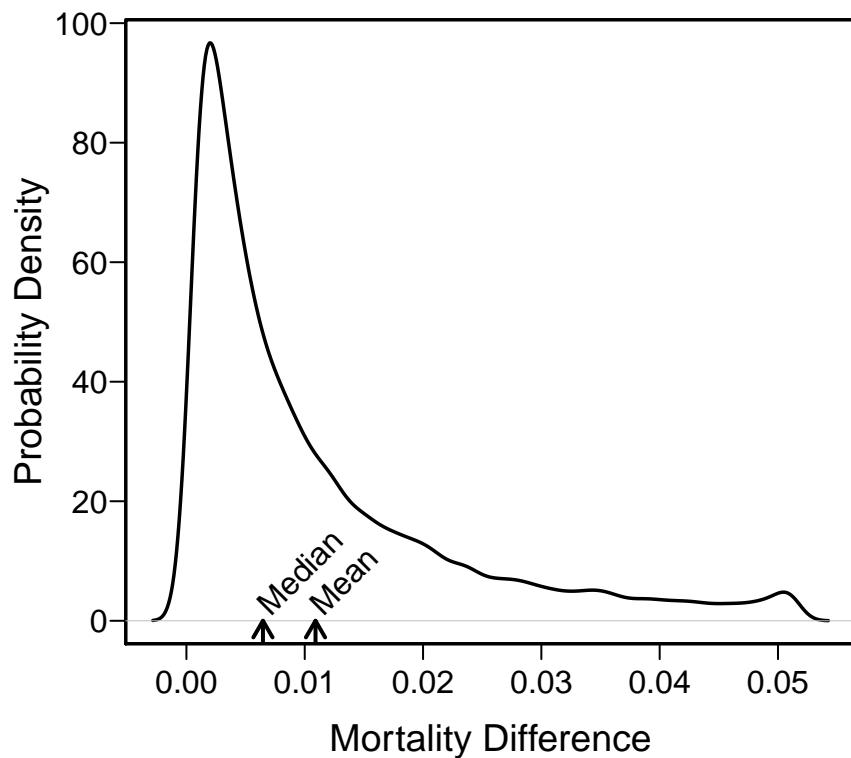


Figure 13.5: Distribution of absolute risk reduction with *t*-PA vs. SK

- Overall mortality difference of 0.011 dominated by high-risk patients

```
load('gusto.rda')
require(rms)
dd <- datadist(gusto); options(datadist='dd')
f <- lrm(day30 ~ tx + age * Killip + pmin(sysbp, 120) +
           lsp(pulse, 50) + pmi + miloc, data=gusto)
cat('{\\smaller ')
```

f

Logistic Regression Model

```
lrm(formula = day30 ~ tx + age * Killip + pmin(sysbp, 120) +
    lsp(pulse, 50) + pmi + miloc, data = gusto)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	40830	LR	χ^2	4173.41	R^2	0.245	C 0.821
0	37979	d.f.		15	g	1.490	D_{xy} 0.642
1	2851	$\Pr(> \chi^2) < 0.0001$			g_r	4.437	γ 0.642
max $ \frac{\partial \log L}{\partial \beta} 5 \times 10^{-6}$					g_p	0.083	τ_a 0.083
				Brier 0.055			

	$\hat{\beta}$	S.E.	Wald Z	Pr(> Z)
Intercept	-3.9541	0.7135	-5.54	<0.0001
tx=SK	0.0738	0.0512	1.44	0.1499
tx=tPA	-0.1338	0.0608	-2.20	0.0276
age	0.0867	0.0026	33.67	<0.0001
Killip=II	2.1146	0.3610	5.86	<0.0001
Killip=III	3.7596	0.7310	5.14	<0.0001
Killip=IV	4.0790	0.8259	4.94	<0.0001
sysbp	-0.0386	0.0017	-23.10	<0.0001
pulse	-0.0221	0.0141	-1.57	0.1168
pulse'	0.0416	0.0143	2.90	0.0037
pmi=yes	0.4664	0.0485	9.62	<0.0001
miloc=Other	0.3048	0.1163	2.62	0.0087
miloc=Anterior	0.5370	0.0443	12.12	<0.0001
age × Killip=II	-0.0216	0.0051	-4.22	<0.0001
age × Killip=III	-0.0363	0.0103	-3.51	0.0004
age × Killip=IV	-0.0323	0.0124	-2.61	0.0090

```
anova(f)
```

Wald Statistics for day30

	χ^2	d.f.	P
tx	15.46	2	0.0004
age (Factor+Higher Order Factors)	1390.07	4	<0.0001
All Interactions	31.13	3	<0.0001
Killip (Factor+Higher Order Factors)	427.94	6	<0.0001
All Interactions	31.13	3	<0.0001
sysbp	533.64	1	<0.0001
pulse	325.19	2	<0.0001
Nonlinear	8.43	1	0.0037
pmi	92.55	1	<0.0001
miloc	146.92	2	<0.0001
age × Killip (Factor+Higher Order Factors)	31.13	3	<0.0001
TOTAL NONLINEAR + INTERACTION	39.17	4	<0.0001
TOTAL	3167.41	15	<0.0001

```
cat('}') #
```

A

```

cof ← coef(f) # vector of regression coefficients
# For cof, X*beta without treatment coefficients estimates logit
# for SK+t-PA combination therapy (reference cell). The coefficient for
# SK estimates the difference in logits from combo to SK. The coefficient
# for tPA estimates the difference in tPA from combo. The mortality
# difference of interest is mortality with SK minus mortality with tPA.
mort.sk ← function(x) plogis(x + cof['tx=SK'])
mort.diff ← function(x)
  ifelse(x < 0, mort.sk(x) - plogis(x + cof['tx=tPA']), NA)
# only define when logit < 0 since U-shaped
n ← nomogram(f, fun=list(mort.sk, mort.diff),
  funlabel=c("30-Day Mortality\nFor SK Treatment",
            "Mortality Reduction by t-PA"),
  fun.at=list(c(.001,.005,.01,.05,.1,.2,.5,.7,.9),
             c(.001,.005,.01,.02,.03,.04,.05)),
  pulse=seq(0,260,by=10), omit='tx', lp=FALSE)
plot(n, varname.label.sep=' ', xfrac=.27, lmgp=.2, cex.axis=.6)

```

Absolute Benefit on Survival Prob.



B

- Cox PH model
- Modeling can uncover time course of treatment effect
- $X_1 = \text{treatment}$, $A = X_2, \dots, X_p$ adjustment variables
- Survival difference is

$$S(t|X_1 = 1, A) - S(t|X_1 = 0, A)$$

$$= S(t|X_1 = 0, A)^{HR} - S(t|X_1 = 0, A)$$

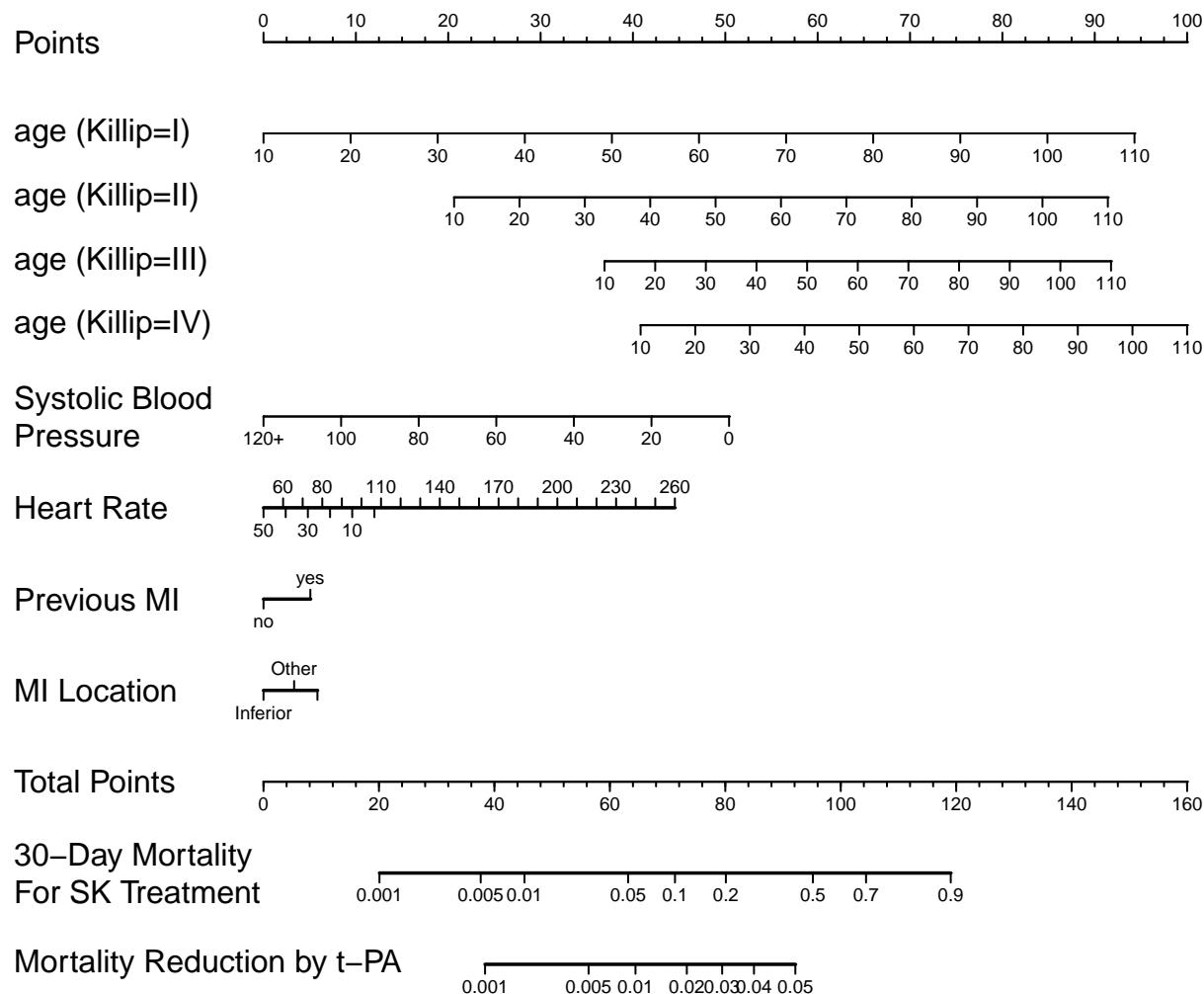


Figure 13.6: Nomogram to predict SK - t -PA mortality difference, based on the difference between two binary logistic models.

```

plot(0, 0, type="n", xlab="Survival for Control Subject",
      ylab="Improvement in Survival",
      xlim=c(0,1), ylim=c(0,.7))      # Fig. 13.7
i ← 0
hr ← seq(.1, .9, by=.1)
for(h in hr) {
  i ← i + 1
  p ← seq(.0001, .9999, length=200)
  p2 ← p ^ h
  d ← p2 - p
  lines(p, d, lty=i)
  maxd ← max(d)
  smax ← p[d==maxd]
  text(smax,maxd+.02, format(h), cex=.6)
}

```

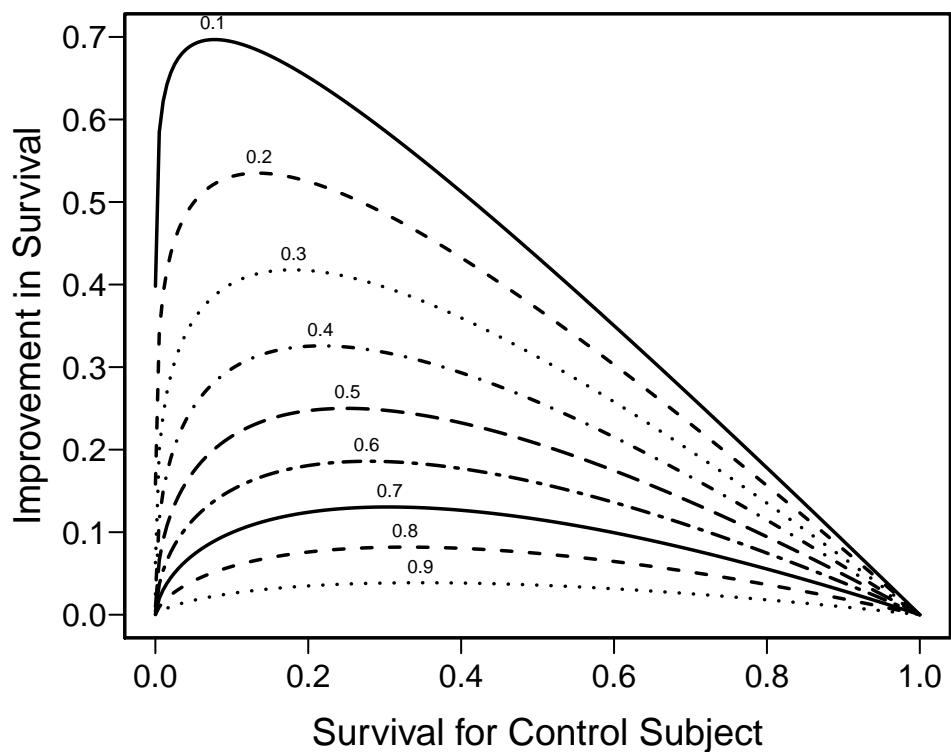
C


Figure 13.7: Relationship between baseline risk, relative treatment effect (hazard ratio — numbers above curves) and absolute treatment effect.

- See also⁵³.

13.7

Cost–Effectiveness Ratios



ancova-ceratio
D

- Effectiveness E (denominator of C–E ratio) is always absolute
- Absolute treatment effectiveness varies greatly with patient characteristics
- → C–E ratio varies greatly
- A C–E ratio based on average E and average C may not apply to any existing patient!
- Need a model to estimate E
- C may also depend on patient characteristics

```
cost.life <- 2400 / delta / 1e6
plot(density(cost.life), xlab='Cost Per Life Saved, $M', main='',
      ylab='Probability Density', xlim=c(0, 6))      # Fig. 13.8
m <- 2400 / mean(delta) / 1e6
u <- par("usr")
arrows(m, u[3], m, 0, length=.1, lwd=2)
text(m,.01,'Cost using\n  average\n    reduction',srt=45,adj=0)
```

E

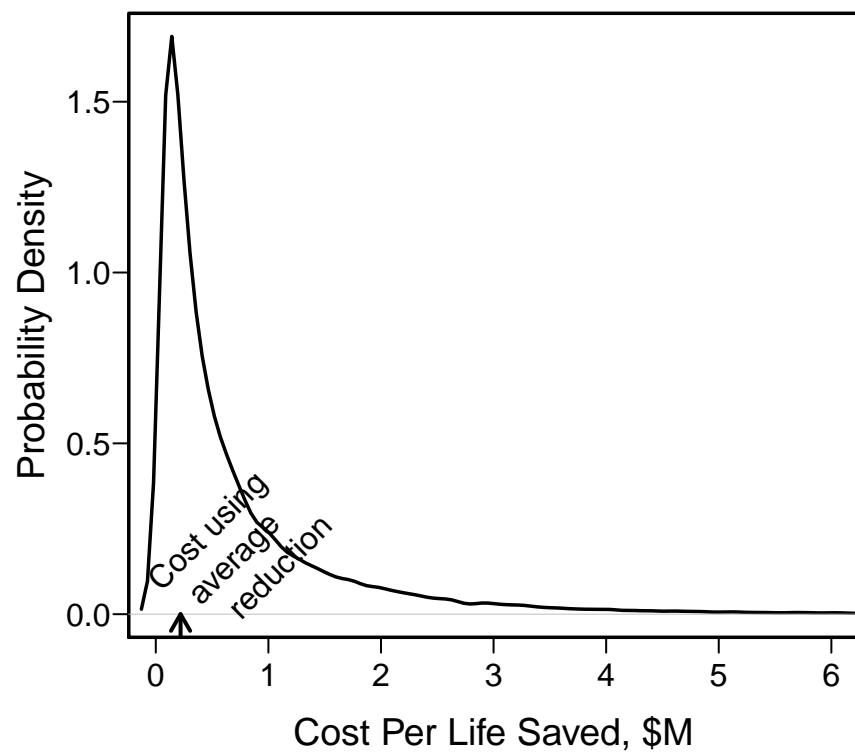


Figure 13.8: Distribution of cost per life saved in GUSTO-I

13.8

Treatment Contrasts for Multi-Site Randomized Trials

ancova-sites
F

- Primary model: covariates, treatment, site main effects
- Planned secondary model to assess consistency of treatment effects over sites (add site \times treatment interactions)
- Advantages for considering sites as random effects (or use penalized MLE to shrink site effects, especially for small sites). See⁷ for a random effects Cox model and a demonstration that treatment effects may be inconsistent when non-zero site main effects are ignored in the Cox model. See also¹¹⁵.
- Types of tests / contrasts when interactions are included⁹⁰:
G

- Type I: not adjusted for center
- Type II: average treatment effect, weighted by size of sites

R rms package command:

```
sites ← levels(site)
contrast(fit, list(treat='b', site=sites),
         list(treat='a', site=sites),
         type='average', weights=table(site))
```

- Type III: average treatment effect, unweighted

```
contrast(fit, list(treat='b', site=sites),
         list(treat='a', site=sites), type='average')
```

Low precision; studies are not powered for Type III tests.

- Another interesting test: combined effects of treatment and site \times treatment interaction; tests whether treatment was effective at *any* site.

13.9

Statistical Plan for Randomized Trials

ancova-plan
H

- When a relevant dataset is available before the trial begins, develop the model from the dataset and use the predicted value as a single adjustment covariate in the trial (Knaus et al.⁵⁴)
- Otherwise: CPMP Working Party: Finalize choice of model, transformations, interactions before merging treatment assignment into analysis dataset.
 Edwards²⁶: Pre-specify family of models that will be used, along with the strategy for selecting the particular model.
 Masked model derivation does not bias treatment effect.
- CPMP guidance²¹
 - "Stratification may be used to ensure balance of treatments across covariates; it may also be used for administrative reasons. The factors that are the basis of stratification should normally be included as covariates in the primary model."
 - Variables known a priori to be strongly, or at least moderately, associated with the primary outcome and/or variables for which there is a strong clinical rationale for such an association should also be considered as covariates in the primary analysis. The variables selected on this basis should be pre-specified in the protocol or the statistical analysis plan.
 - Baseline imbalance observed post hoc should not be considered an appropriate reason for including a variable as a covariate in the primary analysis.
 - Variables measured after randomization and so potentially affected by the treatment should not normally be included as covariates in the primary analysis.
 - If a baseline value of a continuous outcome measure is available, then this should usually be included as a covariate. This applies whether the primary outcome variable is defined as the 'raw outcome' or as the 'change from baseline'.
 - Only a few covariates should be included in a primary analysis. Although larger data sets may support more covariates than smaller ones, justification for including each of the covariates should be provided. (???)
 - In the absence of prior knowledge, a simple functional form (usually either linearity or dichotomising a continuous scale) should be assumed for the relationship between a continuous covariate and the outcome variable. (???)
 - The validity of the model assumptions must be checked when assessing the results. This is particularly important for generalized linear or non-linear models where mis-

specification could lead to incorrect estimates of the treatment effect. Even under ordinary linear models, some attention should be paid to the possible influence of extreme outlying values.

- Whenever adjusted analyses are presented, results of the treatment effect in subgroups formed by the covariates (appropriately categorised, if relevant) should be presented to enable an assessment of the validity of the model assumptions. (???)
- Sensitivity analyses should be pre-planned and presented to investigate the robustness of the primary results. Discrepancies should be discussed and explained. In the presence of important differences that cannot be logically explained—for example, between the results of adjusted and unadjusted analyses—the interpretation of the trial could be seriously affected.
- The primary model should not include treatment by covariate interactions. If substantial interactions are expected a priori, the trial should be designed to allow separate estimates of the treatment effects in specific subgroups.
- Exploratory analyses may be carried out to improve the understanding of covariates not included in the primary analysis, and to help the sponsor with the ongoing development of the drug.
- A primary analysis, unambiguously pre-specified in the protocol or statistical analysis plan, correctly carried out and interpreted, should support the conclusions which are drawn from the trial. Since there may be a number of alternative valid analyses, results based on pre-specified analyses will carry most credibility.”

In confirmatory trials, a model is pre-specified, and it is necessary to pretend that it is true. In most other statistical applications, the choice of model is data-driven, but it is necessary to pretend that it is not.

[Edwards \[26\]](#)

See also Siqueira and Taylor⁹⁷.

- Choose predictors based on expert opinion |
- Impute missing values rather than discarding observations
- Keep all pre-specified predictors in model, regardless of P -value
- Use shrinkage (penalized maximum likelihood estimation) to avoid over-adjustment
- Some guidance for handling missing baseline data in RCTs is in White & Thompson¹¹¹

13.9.1

Sites vs. Covariates

ancova-sites
J

- Site effects (main or interaction) are almost always trivial in comparison with patient-specific covariate effects
- It is not sensible to include site in the model when important covariates are omitted
- The most logical and usually the most strong interaction with treatment is not site but is the severity of disease being treated

13.9.2

Covariate Adjustment vs. Allocation Based on Covariates

K
ancova-plan

The decision to fit prognostic factors has a far more dramatic effect on the precision of our inferences than the choice of an allocation based on covariates or randomization approach and one of my chief objections to the allocation based on covariates approach is that trialists have tended to use the fact that they have balanced as an excuse for not fitting. This is a grave mistake.

Senn [91], p. 3748; see also Senn, Anisimov, and Fedorov [94]

My view ... was that the form of analysis envisaged (that is to say, which factors and covariates should be fitted) justified the allocation and *not vice versa*.

Senn [91], p. 3747

13.10

Summary

The point of view is sometimes defended that analyses that ignore covariates are superior because they are simpler. I do not accept this. A value of $\pi = 3$ is a simple one and accurate to one significant figure . . . However very few would seriously maintain that if should generally be adopted by engineers.

Senn [91], p. 3741

L

- As opposed to simple treatment group comparisons, modeling can
 - Improve precision (linear, log-linear models)
 - Get the “right” treatment effect (nonlinear models)
 - Improve power (almost all models)
 - Uncover outcome patterns, shapes of effects
 - Test/estimate differential treatment benefit
 - Determine whether some patients are too sick or too well to benefit
 - Estimate absolute clinical benefit as a function of severity of illness
 - Estimate meaningful cost-effectiveness ratios
 - Help design the next clinical trial (optimize risk distribution for maximum power)
- Modeling strategy must be well thought-out
 - Not “data mining”
 - Not done to optimize the treatment P -value

13.11

Notes

From a posting by Harrell to the Medstats google group on 19Jan09: I think it is most important to decide what it is you want to estimate, and then formulate a model that will accomplish that. Unlike ordinary linear models, which provide unbiased treatment effects if balanced covariates are mistakenly omitted from the model in an RCT, most models (such as the Cox PH model) result in biased treatment effects even when there is perfect balance in covariates, if the covariates have nonzero effects on the outcome. This is another way of talking about residual outcome heterogeneity.

If you want to estimate the effect of variable X on survival time, averaging over males and females in some strange undocumented way, you can get the population averaged effect of X without including sex in the model. Recognize however this is like comparing some of the males with some of the females when estimating the X effect. This is seldom of interest. More likely we want to know the effect of X for males, the effect for females, and if there is no interaction we pool the two to more precisely estimate the effect of X conditional on sex.

Another way to view this is that the PH assumption is more likely to hold when you condition on covariates than when you don't. No matter what happens though, if PH holds for one case, it cannot hold for the other, e.g., if PH holds after conditioning, it cannot hold when just looking at the marginal effect of X .

See the excellent Tweetorial by Darren Dahly [here](#) and [this](#) supplement to it by Ben Andrew.

[This article](#) by Stephen Senn is also very helpful.

Chapter 14

Transformations, Measuring Change, and Regression to the Mean

14.1

Transformations

- Normality assumption will not always hold
 - Skewed distribution
 - Different standard deviation in different groups
- Transforming data can be a simple method to overcome these problems
- Non-parametric methods (Wilcoxon signed rank, Wilcoxon rank sum) another good option

14.2

Logarithmic Transformation

- Replace individual value by its logarithm

– $u = \log(x)$

- In statistics, always use the *natural* logarithm (base e ; $\ln(x)$)

- Algebra reminders

– $\log(ab) = \log(a) + \log(b)$

– $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

– Inverse of the log function is $\exp(u) = x$, where $\exp(u) = e^u$ and e is a constant ($e = 2.718282\dots$)

14.2.1

Example Dataset

- From Essential Medical Statistics, 13.2 (pre data only)
- Response: Urinary β -thromboglobulin (β -TG) excretion in 24 subjects
- 24 total subjects: 12 diabetic, 12 normal

```
d ← rbind(
  data.frame(status='normal',
             btg=c(4.1, 6.3, 7.8, 8.5, 8.9, 10.4, 11.5, 12.0, 13.8,
                   17.6, 24.3, 37.2)),
  data.frame(status='diabetic',
             btg=c(11.5, 12.1, 16.1, 17.8, 24.0, 28.8, 33.9, 40.7,
                   51.3, 56.2, 61.7, 69.2)))
require(ggplot2)
require(data.table)
```

```
d ← data.table(d)
meds ← d[, j=list(btg = median(btg)), by = status]
p1 ←
  ggplot(d, aes(x=status, y=btg)) +      # Fig. 14.1
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
```

```

geom_errorbar(aes(ymin=..y.., ymax=..y..), width=.25, size=1.3, data=meds) +
  xlab('') + ylab(expression(paste(beta-TG, ', (ng/day/100 ml creatinine)')))) +
  coord_flip()
p2 <- ggplot(d, aes(x=status, y=btg)) +
  scale_y_log10(breaks=c(4,5,10,15,20,30,40,60,80)) +
  geom_dotplot(binaxis='y', stackdir='center', position='dodge') +
  xlab('') + ylab(expression(paste(beta-TG, ', (ng/day/100 ml creatinine)')))) +
  coord_flip()
arrGrob(p1, p2, ncol=2)

```

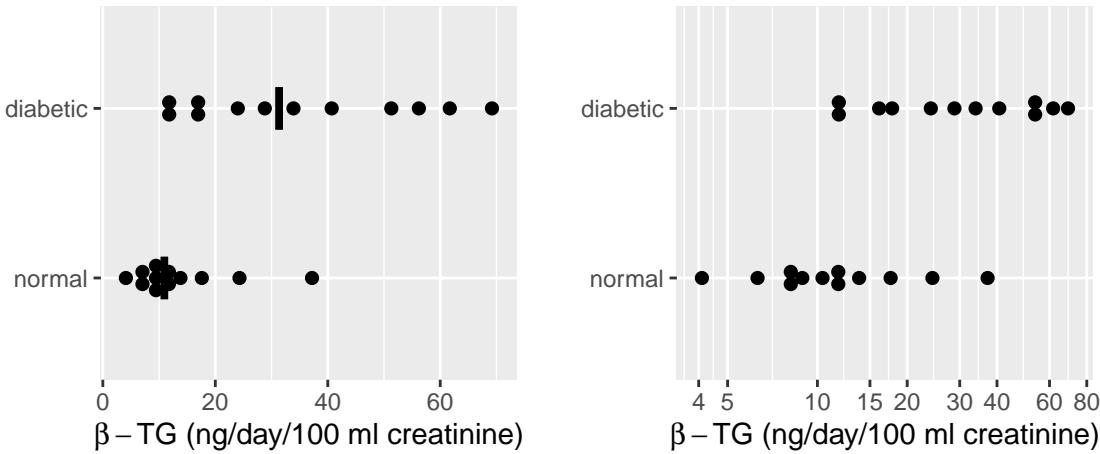


Figure 14.1: β -TG levels by diabetic status with a median line. The left plot is on the original (non-transformed) scale and includes median lines. The right plot displays the data on a log scale.

- Original scale
 - Normal: $\bar{x}_1 = 13.53$, $s_1 = 9.194$, $n_1 = 12$
 - Diabetic: $\bar{x}_2 = 35.28$, $s_2 = 20.27$, $n_2 = 12$

- Logarithm scale
 - Normal: $\bar{x}_1^* = 2.433$, $s_1^* = 0.595$, $n_1 = 12$
 - Diabetic: $\bar{x}_2^* = 3.391$, $s_2^* = 0.637$, $n_2 = 12$

- t -test on log-transformed data
 - $s_{pool} = \sqrt{\frac{11 \times .595^2 + 11 \times .637^2}{22}} = 0.616$
 - $t = \frac{2.433 - 3.391}{0.616 \sqrt{1/12 + 1/12}} = -3.81$, $df = 22$, $p = 0.001$

- Confidence Intervals (0.95 CI)

- Note that $t_{.975,22} = 2.074$
- For Normal subjects, a CI for the mean $\log \beta\text{-TG}$ is

$$\begin{aligned} 0.95 \text{ CI} &= 2.433 - 2.074 \times \frac{0.595}{\sqrt{12}} \text{ to } 2.433 + 2.074 \frac{0.595}{\sqrt{12}} \\ &= 2.08 \text{ to } 2.79 \end{aligned}$$

- Can transform back to original scale by using the antilog function $e(u)$ to estimate **medians**

$$\begin{aligned} \text{Geometric mean} &= e^{2.433} = 11.39 \\ 0.95 \text{ CI} &= e^{2.08} \text{ to } e^{2.79} \\ &= 7.98 \text{ to } 16.27 \end{aligned}$$

```
t.test(btg ~ status, data=d)
```

```
Welch Two Sample t-test

data: btg by status
t = -3.3838, df = 15.343, p-value = 0.003982
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-35.41024 -8.07309
sample estimates:
mean in group normal mean in group diabetic
13.53333            35.27500
```

```
t.test(log(btg) ~ status, data=d)
```

```
Welch Two Sample t-test

data: log(btg) by status
t = -3.8041, df = 21.9, p-value = 0.0009776
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.4792986 -0.4352589
sample estimates:
mean in group normal mean in group diabetic
2.433349            3.390628
```

- Could also use a non-parametric test (Wilcoxon rank sum)

```
wilcox.test(btg ~ status, data=d)
```

```
Wilcoxon rank sum test with continuity correction
```

```
data: btg by status  
W = 18.5, p-value = 0.002209  
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(log(btg) ~ status, data=d)
```

```
Wilcoxon rank sum test with continuity correction  
  
data: log(btg) by status  
W = 18.5, p-value = 0.002209  
alternative hypothesis: true location shift is not equal to 0
```

- Note that non-parametric test is the same for the log-transformed outcomes

14.2.2

Limitations of log transformations

- Can only be used on positive numbers
 - Sometimes use $u = \log(x + 1)$
- Is very arbitrary to the choice of the origin
- Not always useful or the best transformation
- Sometimes use a dimensionality argument, e.g., take cube root of volume measurements or per unit of volume counts like blood counts
- Cube and square roots are fine with zeros

14.3

Analysis of Paired Observations

- Frequently one makes multiple observations on same experimental unit
- Can't analyze as if independent
- When two observations made on each unit (e.g., pre–post), it is common to summarize each pair using a measure of effect → analyze effects as if (unpaired) raw data
- Most common: simple difference, ratio, percent change
- Can't take effect measure for granted
- Subjects having large initial values may have largest differences
- Subjects having very small initial values may have largest post/pre ratios

14.4

What's Wrong with Change in General?

14.4.1

Change from Baseline in Randomized Studies

Many authors and pharmaceutical clinical trialists make the mistake of analyzing change from baseline instead of making the raw follow-up measurements the primary outcomes, covariate-adjusted for baseline. The purpose of a parallel-group randomized clinical trial is to compare the parallel groups, not to compare a patient with herself at baseline. The central question is for two patients with the same pre measurement value of x , one given treatment A and the other treatment B, will the patients tend to have different post-treatment values? This is exactly what analysis of covariance assesses. Within-patient change is affected strongly by regression to the mean and measurement error. When the baseline value is one of the patient inclusion/exclusion criteria, the only meaningful change score requires one to have a second baseline measurement post patient qualification to cancel out much of the regression to the mean effect. It is the second baseline that would be subtracted from the follow-up measurement.

The savvy researcher knows that analysis of covariance is required to “rescue” a change score analysis. This effectively cancels out the change score and gives the right answer even if the slope of post on pre is not 1.0. But this works only in the linear model case, and it can be confusing to have the pre variable on both the left and right hand sides of the statistical model. And if Y is ordinal but not interval-scaled, the difference in two ordinal variables is no longer even ordinal. Think of how meaningless difference from baseline in ordinal pain categories are. A major problem in the use of change score summaries, even when a correct analysis of covariance has been done, is that many papers and drug product labels still quote change scores out of context. Patient-reported outcome scales with floor or ceiling effects are particularly problematic. Analysis of change loses the opportunity to do a robust, powerful analysis using a covariate-adjusted ordinal response model such as the proportional odds or proportional hazards model. Such ordinal response models do not require one to be correct in how to transform Y .

Not only is it very problematic to analyze change scores in parallel group designs; it is problematic to even compute them. To compute change scores requires many

assumptions to hold, e.g.:

1. the variable is not used as an inclusion/exclusion criterion for the study, otherwise regression to the mean will be strong
2. if the variable is used to select patients for the study, a second post-enrollment baseline is measured and this baseline is the one used for all subsequent analysis
3. the post value must be linearly related to the pre value
4. the variable must be perfectly transformed so that subtraction "works" and the result is not baseline-dependent
5. the variable must not have floor and ceiling effects
6. the variable must have a smooth distribution
7. the slope of the pre value vs. the follow-up measurement must be close to 1.0 when both variables are properly transformed (using the same transformation on both)

Details about problems with analyzing change may be found [here](#), and references may be found [here](#).

Regarding 3. above, if pre is not linearly related to post, there is no transformation that can make a change score work. Two unpublished examples illustrate this problem. In a large depression study using longitudinal measurements of the Hamilton-D depression scale, there was a strong nonlinear relationship between baseline Hamilton D and the final measurement. The form of the relationship was a flattening of the effect for larger Ham D, indicating there are patients with severe depression at baseline who can achieve much larger reductions in depression than an average change score would represent (and likewise, patients with mild to moderate depression at baseline would have a reduction in depression symptoms that is far less than the average change indicates). Doing an ordinal ANCOVA adjusting for a smooth nonlinear effect of baseline using a spline function would have been a much better analysis than the change from baseline analysis done by study leaders. In the second example, a similar result was found for a quality of life measure, KCCQ. Again, a flattening relationship for large KCCQ indicated that subtracting from baseline provided a nearly meaningless average change score.

Regarding 7. above, often the baseline is not as relevant as thought and the slope will be less than 1. When the treatment can cure every patient, the slope will be zero. When the baseline variable is irrelevant, ANCOVA will estimate a slope of approximately zero and will effectively ignore the baseline. Change will baseline will make the change

score more noisy than just analyzing the final raw measurement. In general, when the relationship between pre and post is linear, the correlation between the two must exceed 0.5 for the change score to have lower variance than the raw measurement.

Bland and Altman⁹ have an excellent [article](#) about how misleading change from baseline is for clinical trials. [This](#) blog article has several examples of problematic change score analyses in the clinical trials literature.

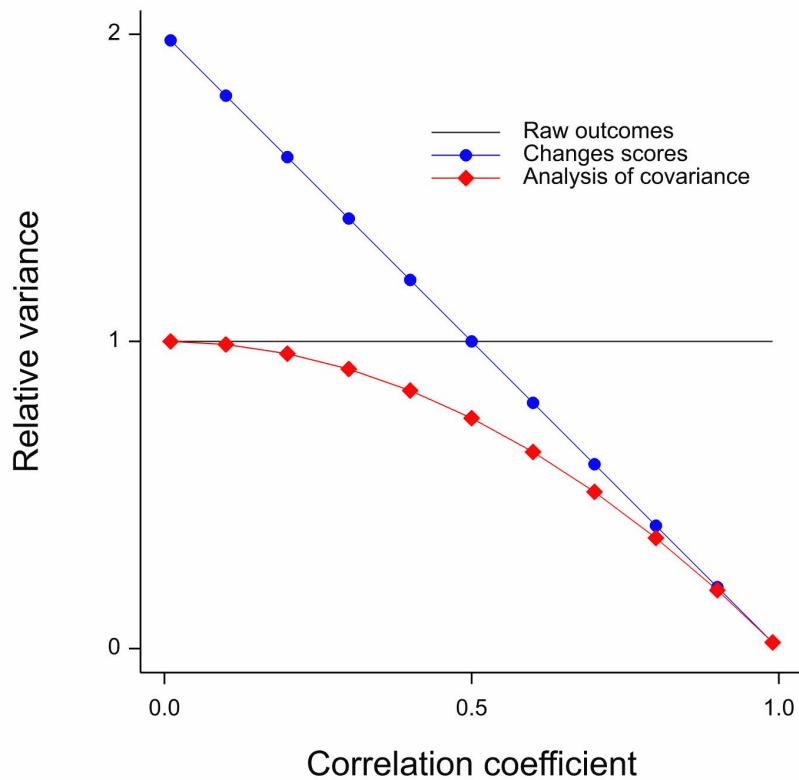
The following R code exemplifies how to do a powerful, robust analysis without computing change from baseline. This analysis addresses the fundamental treatment question posed at the top of this section. This example uses a semiparametric ANCOVA that utilizes only the ranks of Y and that provides the same test statistic no matter how Y is transformed. It uses the proportional odds model, with no binning of Y , using the `rms` package `orm` function^a. This is a generalization of the Wilcoxon test—it would be almost identical to the Wilcoxon test had the baseline effect been exactly flat. Note that the proportional odds assumption is more likely to be satisfied than the normal residual, equal variance assumption of the ordinary linear model.

```
require(rms)
# Fit a smooth flexible relationship with baseline value y0
# (restricted cubic spline function with 4 default knots)
f <- orm(y ~ rcs(y0, 4) + treatment)
f
anova(f)
# Estimate mean y as a function of y0 and treatment
M <- Mean(f) # creates R function to compute the mean
plot(Predict(f, treatment, fun=M)) # y0 set to median since not varied
# To allow treatment to interact with baseline value in a general way:
f <- orm(y ~ rcs(y0, 4) * treatment)
plot(Predict(f, y0, treatment)) # plot y0 on x-axis, 2 curves for 2 treatments
# The above plotted the linear predictor (log odds); can also plot mean
```

Kristoffer Magnusson has an excellent paper [Change over time is not “treatment response”](#)

Stephen Senn⁹² Chapter 7 has a very useful graphical summary of the large-sample variances (inefficiency) of change scores vs. ANCOVA vs. ignoring baseline completely, as a function of the (assumed linear) correlation between baseline and outcome.

^a`orm` can easily handle thousands of intercepts (thousands of distinct Y values).



14.4.2

Special Problems Caused By Non-Monotonic Relationship With Ultimate Outcome

Besides the strong assumptions made about how variables are transformed before a difference is calculated, there are many types of variables for which change can never be interpreted without reference to the starting point, because the relationship between the variable in question and an ultimate outcome is not even monotonic. A good example is the improper definitions of acute kidney injury (AKI) that have been accepted without question. Many of the definitions are based on change in serum creatinine (SCr). Problems with the definitions include

1. The non-monotonic relationship between SCr and mortality demonstrates that it is not healthy to have very low SCr. This implies that increases in SCr for very low starting SCr may not be harmful as assumed in definitions of AKI.

2. Given an earlier SCr measurement and a current SCr, the earlier measurement is not very important in predicting mortality once one adjusts for the last measurement. Hence a change in SCr is not very predictive and the current SCr is all-important.

As an example consider the estimated relationship between baseline SCr and mortality in critically ill ICU patients.

```
require(rms)
```

```
load('~/Analyses/SUPPORT/combined.sav')
combined <- subset(combined,
  select=c(id, death, d.time, hospdead, dzgroup, age, raceh, sex))
load('~/Analyses/SUPPORT/combphys.sav')
combphys <- subset(combphys, !is.na(crea1+crea3),
  select=c(id, crea1, crea3, crea7, crea14, crea25, alb3,
  meanbp3, pafi3, wblc3))
w <- merge(combined, combphys, by='id')
u <- 'mg/dl'
w <- upData(w, labels=c(crea1='Serum Creatinine, Day 1',
  crea3='Serum Creatinine Day 3',
  crea14='Serum Creatinine Day 14'),
  units=c(crea1=u, crea3=u, crea7=u, crea14=u, crea25=u))
```

Input object size:	1739440 bytes;	17 variables	10279 observations
New object size:	1740560 bytes;	17 variables	10279 observations

```
w <- subset(w, crea1 < 2)
dd <- datadist(w); options(datadist='dd')

h <- lrm(hospdead ~ rcs(crea1, 5) + rcs(crea3, 5), data=w)
anova(h) #
```

	Wald Statistics		Response: hospdead
Factor	Chi-Square	d.f.	P
crea1	19.52	4	0.0006
Nonlinear	15.60	3	0.0014
crea3	108.11	4	<.0001
Nonlinear	49.98	3	<.0001
TOTAL NONLINEAR	132.06	6	<.0001
TOTAL	217.11	8	<.0001

```
h <- lrm(hospdead ~ sex * rcs(crea3, 5), data=w)
p <- Predict(h, crea3, sex, fun=plogis)
ggplot(p, ylab='Risk of Hospital Death') # Fig. 14.2
```

We see that the relationship is very non-monotonic so that it is impossible for change in SCr to be relevant by itself unless the study excludes all patients with SCr < 1.05. To put this in perspective, in the NHANES study of asymptomatic subjects, a very significant proportion of subjects have SCr < 1.

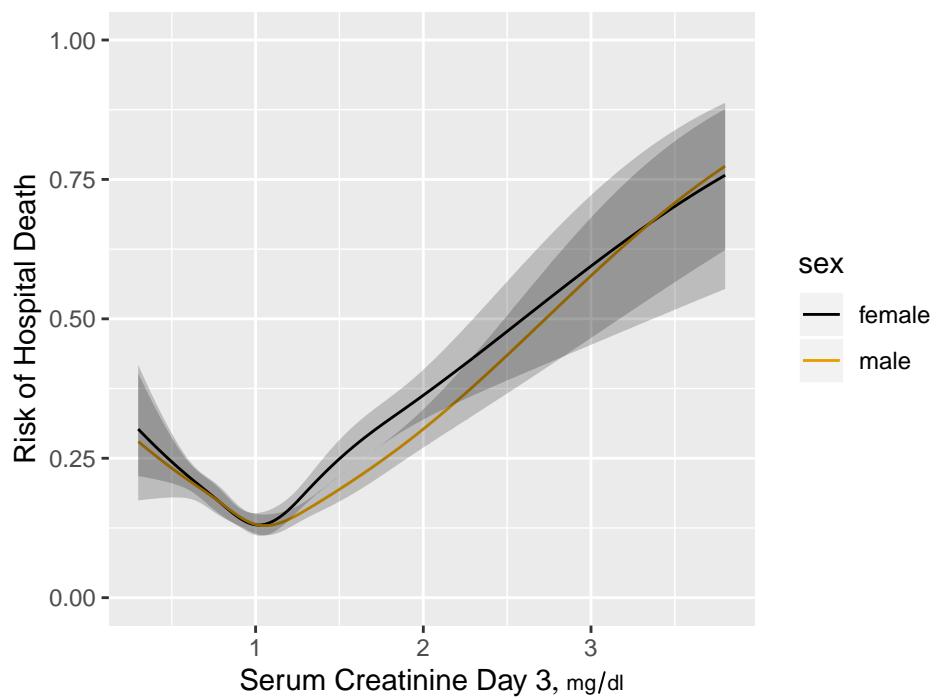


Figure 14.2: Estimated risk of hospital death as a function of day 3 serum creatinine and sex for 7772 critically ill ICU patients having day 1 serum creatinine < 2 and surviving to the start of day 3 in the ICU

14.5

What's Wrong with Percent Change?

- Definition

$$\% \text{ change} = \frac{\text{first value} - \text{second value}}{\text{second value}} \times 100$$

- The first value is often called the new value and the second value is called the old value, but this does not fit all situations

- Example

- Treatment A: 0.05 proportion having stroke

- Treatment B: 0.09 proportion having stroke

- The point of reference (which term is used in the denominator?) will impact the answer

- Treatment A reduced proportion of stroke by 44%

- Treatment B increased proportion by 80%

- Two increases of 50% result in a total increase of 125%, not 100%

- Math details: If x is your original amount, two increases of 50% is $x * 1.5 * 1.5$.

- Then, $\% \text{ change} = (1.5 * 1.5 * x - x) / x = x * (1.5 * 1.5 - 1) / x = 1.25$, or a 125% increase

- Percent change (or ratio) not a symmetric measure

- A 50% increase followed by a 50% decrease results in an overall decrease (not no change)

- * Example: 2 to 3 to 1.5

- A 50% decrease followed by a 50% increase results in an overall decrease (not no change)

- * Example: 2 to 1 to 1.5

- Unless percents represent proportions times 100, it is not appropriate to compute descriptive statistics (especially the mean) on percents.
 - For example, the correct summary of a 100% increase and a 50% decrease, if they both started at the same point, would be 0% (not 25%).
- Simple difference or log ratio are symmetric

14.6

Objective Method for Choosing Effect Measure

- Goal: Measure of effect should be as independent of baseline value as possible^b
- Plot difference in pre and post values vs. the pre values. If this shows no trend, the simple differences are adequate summaries of the effects, i.e., they are independent of initial measurements.
- If a systematic pattern is observed, consider repeating the previous step after taking logs of both the pre and post values. If this removes any systematic relationship between the baseline and the difference in logs, summarize the data using logs, i.e., take the effect measure as the log ratio.
- Other transformations may also need to be examined

^bBecause of regression to the mean, it may be impossible to make the measure of change truly independent of the initial value. A high initial value may be that way because of measurement error. The high value will cause the change to be less than it would have been had the initial value been measured without error. Plotting differences against averages rather than against initial values will help reduce the effect of regression to the mean.

14.7**Example Analysis: Paired Observations****14.7.1****Dataset description**

- Dataset is an extension of the diabetes dataset used earlier in this chapter
- Response: Urinary β -thromboglobulin (β -TG) excretion in 24 subjects
- 24 total subjects: 12 diabetic, 12 normal
- Add a “post” measurement (previous data considered the “pre” measurement)

14.7.2**Example Analysis**

```
# Now add simulated some post data to the analysis of beta TG data
# Assume that the intervention effect (pre -> post effect) is
# multiplicative (x 1/4) and that there is a multiplicative error
# in the post measurements
set.seed(13)
d$pre  ← d$btg
d$post  ← exp(log(d$pre) + log(.25) + rnorm(24, 0, .5))
# Make plots on the original and log scales
p1  ← ggplot(d, aes(x=pre, y=post - pre, color=status)) +
  geom_point() + geom_smooth() + theme(legend.position='bottom')
# Use problematic asymmetric % change
p2  ← ggplot(d, aes(x=pre, y=100*(post - pre)/pre,
  color=status)) + geom_point() + geom_smooth() +
  xlab('pre') + theme(legend.position='none') +
  ylim(-125, 0)
p3  ← ggplot(d, aes(x=pre, y=log(post / pre),
  color=status)) + geom_point() + geom_smooth() +
  xlab('pre') + theme(legend.position='none') + ylim(-2.5, 0)
arrGrob(p1, p2, p3, ncol=2) # Fig. 14.3
```

```
with(d, {
  print(t.test(post - pre))
  print(t.test(100*(post - pre) / pre)) # improper
  print(t.test(log(post / pre)))
  print(wilcox.test(post - pre))
  print(wilcox.test(100*(post - pre) / pre)) # improper
  print(wilcox.test(log(post / pre)))
```

{ }

One Sample t-test

```
data: post - pre
t = -5.9366, df = 23, p-value = 4.723e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-22.68768 -10.96215
sample estimates:
mean of x
-16.82492
```

One Sample t-test

```
data: 100 * (post - pre)/pre
t = -23.864, df = 23, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-75.19217 -63.19607
sample estimates:
mean of x
-69.19412
```

One Sample t-test

```
data: log(post/pre)
t = -13.147, df = 23, p-value = 3.5e-12
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-1.483541 -1.080148
sample estimates:
mean of x
-1.281845
```

Wilcoxon signed rank test

```
data: post - pre
V = 0, p-value = 1.192e-07
alternative hypothesis: true location is not equal to 0
```

Wilcoxon signed rank test

```
data: 100 * (post - pre)/pre
V = 0, p-value = 1.192e-07
alternative hypothesis: true location is not equal to 0
```

Wilcoxon signed rank test

```
data: log(post/pre)
V = 0, p-value = 1.192e-07
```

```
alternative hypothesis: true location is not equal to 0
```

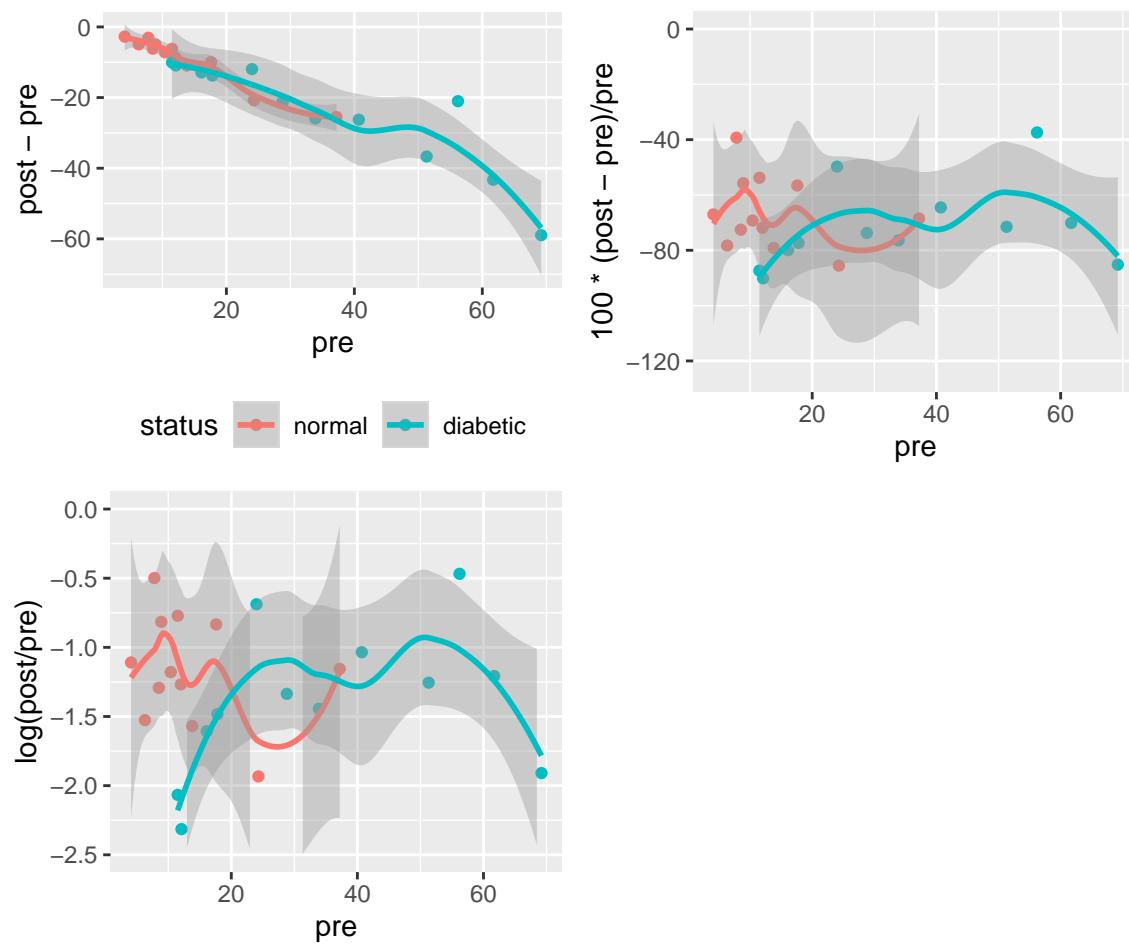


Figure 14.3: Difference vs. baseline plots for three transformations

Note: In general, the three Wilcoxon signed-rank statistics will not agree on each other. They depend on the symmetry of the difference measure.

14.8

Regression to the Mean



- One of the most important of all phenomena regarding data and estimation
- Occurs when subjects are selected because of their values
- Examples:
 1. Intersections with frequent traffic accidents will have fewer accidents in the next observation period if no changes are made to the intersection
 2. The surgeon with the highest operative mortality will have a significant decrease in mortality in the following year
 3. Subjects screened for high cholesterol to qualify for a clinical trial will have lower cholesterol once they are enrolled
- Observations from a randomly chosen subject are unbiased for that subject
- But subjects *selected* because they are running high or low are selected partially because their measurements are atypical for themselves (i.e., selected *because of* measurement error)
- Future measurements will “regress to the mean” because measurement errors are random
- For a classic misattribution of regression to the mean to a treatment effect see [this^c](#).

Classic paper on shrinkage: Efron & Morris²⁷

- Shrinkage is a way of discounting observed variation that accounts for regression to the mean
- In their example you can see that the variation in batting averages for the first 45 at bats is unrealistically large

^cIn their original study, the social workers enrolled patients having 10 or more hospital admissions in the previous year and showed that after their counseling, the number of admissions in the next year was less than 10. The same effect might have been observed had the social workers given the patients horoscopes or weather forecasts. This was reported in an abstract for the AHA meeting that has since been taken down from circ.ahajournals.org.

- Shrunken estimates (middle) have too little variation but this discounting made the estimates closer to the truth (final batting averages at the end of the season)
- You can see the regression to the mean for the *apparently* very hot and very cold hitters

```

nam <- c('Roberto Clemente','Frank Robinson','Frank Howard','Jay Johnstone',
        'Ken Berry','Jim Spencer','Don Kessinger','Luis Alvarado',
        'Ron Santo','Ron Swoboda','Del Unser','Billy Williams',
        'George Scott','Rico Petrocelli','Ellie Rodriguez',
        'Bert Campaneris','Thurman Munson','Max Alvis')
initial <- c(18,17,16,15,14,14,13,12,11,11,10,10,10,10,10,9,8,7)/45
season <- c(345,297,275,220,272,270,265,210,270,230,265,258,306,265,225,
           283,320,200)/1000
initial.shrunk <- c(294,288,280,276,275,275,270,265,262,263,258,256,
                     257,256,257,252,245,240)/1000
plot(0,0,xlim=c(0,1),ylim=c(.15,.40),type='n',axes=F,xlab='',ylab='')
n <- 18
x1 <- .5
x2 <- .75
x3 <- 1
points(rep(x1,n), initial)
points(rep(x2,n), initial.shrunk)
points(rep(x3,n), season)

for(i in 1:n) lines(c(x1,x2,x3),c(initial[i],initial.shrunk[i],season[i]),
                     col=i, lwd=2.5)
axis(2)
par(xpd=NA)
text(c(x1,x2+.01, x2+.25),rep(.12,3),c('First 45 ABs','Shrunken\nEstimates',
      'Rest of\nSeason'))
for(a in unique(initial)) {
  s <- initial==a
  w <- if(sum(s) < 4) paste(nam[s],collapse=', ')
  else {
    j <- (1:n)[s]
    paste(nam[j[1]],', ',nam[j[2]],', ',nam[j[3]],'\n',
          nam[j[4]],', ',nam[j[5]],sep=' ')
  }
  text(x1-.02, a, w, adj=1, cex=.9)
}

```

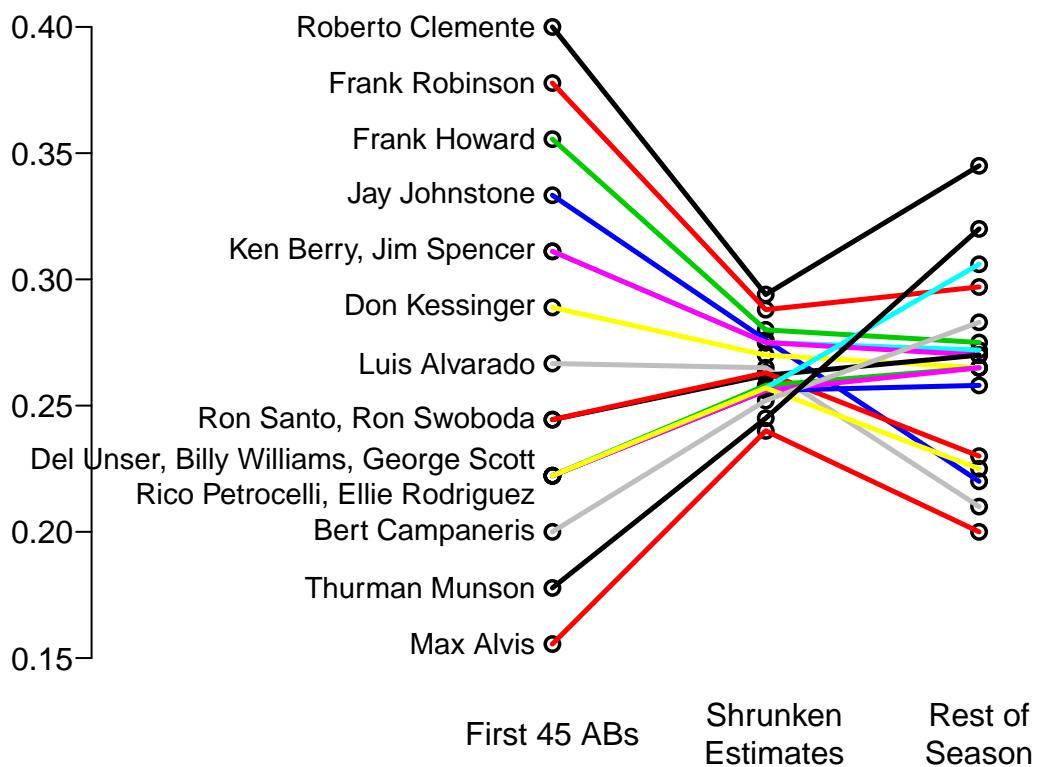


Figure 14.4: Initial batting averages as estimates of final batting averages for players, along with shrunken estimates that account for regression to the mean

Chapter 15

Serial Data

15.1

Introduction



Serial data, also called longitudinal data, repeated measures, or panel data, present a special challenge in that the response variable is multivariate, and that the responses measured at different times (or other conditions such as doses) but on the same subject are correlated with each other. One expects correlations between two measurements measured closer together will exceed correlations for two distant measurements on the same subject. Analyzing the repeated measurements just as though they were independent measurements falsely inflates the sample size and results in a failure to preserve type I error and confidence interval coverage. For example, having three measurements on 10 subjects results in 30 measurements, but depending on the correlations among the three measurements within subject, the effective sample size might be 16, for example. In other words, the statistical information for collecting 3 measurements on each of 10 subjects might provide the same statistical information and power as having one measurement on each of 16 subjects. The statistical information would however be less than that from 30 subjects measured once, if the intra-subject correlation exceeds zero.

The most common problems in analyzing serial data are

B

1. treating repeated measurements per subject as if they were from separate subjects
2. using two-way ANOVA as if different measurement times corresponded to different groups of subjects
3. using repeated measures ANOVA which assumes that the correlation between any

two points within subject is the same regardless of how far apart the measures were timed

4. analyzing more than 3 time points as if time is a categorical rather than a continuous variable

- multiplicity problem
- analyses at different times may be inconsistent since times are not connected
- loss of power by not jointly analyzing the serial data

15.2

Analysis Options

C

There are several overall approaches to the analysis of serial measurements. Some of the older approaches such as multiple t -tests and repeated measures ANOVA are now considered obsolete because of the availability of better methods that are more flexible and robust^a. Separate t -tests at each time point do not make good use of available information, use an inefficient estimate of σ^2 , do not interpolate time points, and have multiple comparison problems. Since a multiplicity adjustment for multiple correlated t -tests is not model-based, the resulting confidence intervals and P -values are conservative. To preserve type I error, one always must sacrifice type II error, but in this case the sacrifice is too severe. In addition, investigators are frequently confused by the t -test being “significant” at one time point and not at another, and make the unwarranted claim that the treatment is effective only at the first time. Besides not recognizing the *absence of evidence is not evidence for absence* problem, the investigator can be misled by increased variability in response at the second time point driving the t ratio towards zero. This variability may not even be real but may reflect instability in estimating σ separately at each time point.

Especially when there are more than three unique measurement times, it is advisable to model time as a continuous variable. When estimation of the time-response profile is of central importance, that may be all that's needed. When comparing time-response profiles (e.g., comparing two treatments) one needs to carefully consider the characteristics of the profiles that should be tested, i.e., where the type I and II errors should be directed, for example

D

- difference in slope
- difference in area under the curve
- difference in mean response at the last planned measurement time
- difference in mean curves at *any* time, i.e., whether the curves have different heights or shapes anywhere

The first 3 are 1 d.f. tests, and the last is a 2 d.f. test if linearity is assumed, and > 2

^aRepeated measures ANOVA makes stringent assumptions and requires complex adjustments for within-subject correlation.

d.f. if fitting a polynomial or spline function.

15.2.1

Joint Multivariate Models

E

This is the formal fully-specified statistical model approach whereby a likelihood function is formed and maximized. This handles highly imbalanced data, e.g., one subject having one measurement and another having 100. It is also the most robust approach to non-random subject dropouts. These two advantages come from the fact that full likelihood models “know” exactly how observations from the same subject are connected to each other.

Examples of full likelihood-based models include generalized least squares, mixed effects models, and Bayesian hierarchical models. Generalized least squares only handles continuous Y and assumes multivariate normality. It does not allow different subjects to have different slopes. But it is easy to specify and interpret, to have its assumptions checked, and it runs faster. Mixed effects models can handle multiple hierarchical levels (e.g., state/hospital/patient) and random slopes whereby subjects can have different trajectories. Mixed effects models can be generalized to binary and other endpoints but lose their full likelihood status somewhat when these extensions are used, unless a Bayesian approach is used.

Generalized least squares, formerly called growth curve models, is the oldest approach and has excellent performance when Y is conditionally normally distributed.

15.2.2

GEE

F

Generalized estimating equations is usually based on a working independence model whereby ordinary univariate regressions are fitted on a combined dataset as if all observations are uncorrelated, and then an after-the-fit correction for intra-cluster correlation is done using the cluster sandwich covariance estimator or the cluster bootstrap. GEE is very non-robust to non-random subject dropout; it assumes missing response values are missing completely at random. It may also require large sample sizes for P -values and confidence intervals to be accurate. An advantage of GEE is that it extends easily to every type of response variable, including binary, ordinal, polytomous, and time-to-event

responses.

15.2.3

Summary Measures



G

H

A simple and frequently effective approach is to summarize the serial measures from each subject using one or two measures, then to analyze these measures using traditional statistical methods that capitalize on the summary measures being independent for different subjects. This has been called

1. Two-stage derived variable analysis²⁴
2. Response feature analysis²⁵
3. Longitudinal analysis through summary measures⁶⁴

An excellent overview may be found in Matthews et al. [64], Dupont [25] (Chapter 11), and Senn, Stevens, and Chaturvedi [95].

Frequently chosen summary measures include the area under the time-response curve, slope, intercept, and consideration of multiple features simultaneously, e.g., intercept, coefficient of time, coefficient of time squared when fitting each subject's data with a quadratic equation. This allows detailed analyses of curve shapes.

15.3

Case Study

15.3.1

Data and Summary Measures

Consider the isoproterenol dose-response analysis of Dupont [25] of the original data from Lang^b. Twenty two normotensive men were studied, 9 of them black and 13 white. Blood flow was measured before the drug was given, and at escalating doses of isoproterenol. Most subjects had 7 measurements, and these are not independent.

```
require(Hmisc)
require(data.table)    # elegant handling of aggregation

require(ggplot2)
d <- csv.get(paste('http://biostat.mc.vanderbilt.edu',
                    'dupontwd/wddtext/data/11.2.Long.Isoproterenol.csv',
                    sep='/'))
d <- upData(d, keep=c('id', 'dose', 'race', 'fbf'),
            race =factor(race, 1:2, c('white', 'black')),
            labels=c(fbf='Forearm Blood Flow'),
            units=c(fbf='ml/min/dl'))
```

Input object size:	12680 bytes;	8 variables	154 observations
Modified variable	race		
Kept variables	id, dose, race, fbf		
New object size:	6216 bytes;	4 variables	154 observations

```
d <- data.table(d)
setkey(d, id, race)
```

```
# Fit subject-by-subject spline fits and either return the coefficients,
# the estimated area under the curve from [0,400], or evaluate each
# subject's fitted curve over a regular grid of 150 doses
# Area under curve is divided by 400 to get a mean function
require(rms)
```

```
g <- function(x, y, what=c('curve', 'coef', 'area')) {
  what <- match.arg(what)    # 'curve' is default
  knots <- c(20, 60, 150)
  f <- ols(y ~ rcs(x, knots))
  xs <- seq(0, 400, length=150)
  switch(what,
    coef = {k <- coef(f)
              list(b0 = k[1], b1=k[2], b2=k[3])},
    curve= {x <- seq(0, 400, length=150)
              list(dose=xs, fbf=predict(f, data.frame(x=xs)))},
```

^bCC Lang et al. NEJM 333:155-60, 1995

```

area = {antiDeriv = rcsplineFunction(knots, coef(f),
                                      type='integral')
        list(dose = 400, fbf=y[x == 400],
             area = antiDeriv(400) / 400,
             tarea = areat(x, y) / 400) }
}

# Function to use trapezoidal rule to compute area under the curve
areat <- function(x, y) {
  i <- ! is.na(x + y)
  x <- x[i]; y <- y[i]
  i <- order(x)
  x <- x[i]; y <- y[i]
  if(! any(x == 400)) NA else
  sum(diff(x) * (y[-1] + y[-length(y)]))/2
}

w <- d[, j=g(dose, fbf), by = list(id, race)] # uses data.table package
a <- d[, j=g(dose, fbf, what='area'), by = list(id, race)]

ggplot(d, aes(x=dose, y=fbf, color=factor(id))) +    # Fig. 15.1 J
  geom_line() + geom_line(data=w, alpha=0.25) +
  geom_text(aes(label = round(area,1)), data=a, size=2.5,
            position=position_dodge(width=50)) +
  xlab('Dose') + ylab(label(d$fbf, units=TRUE, plot=TRUE)) +
  facet_grid(~ race) +
  guides(color=FALSE)

```

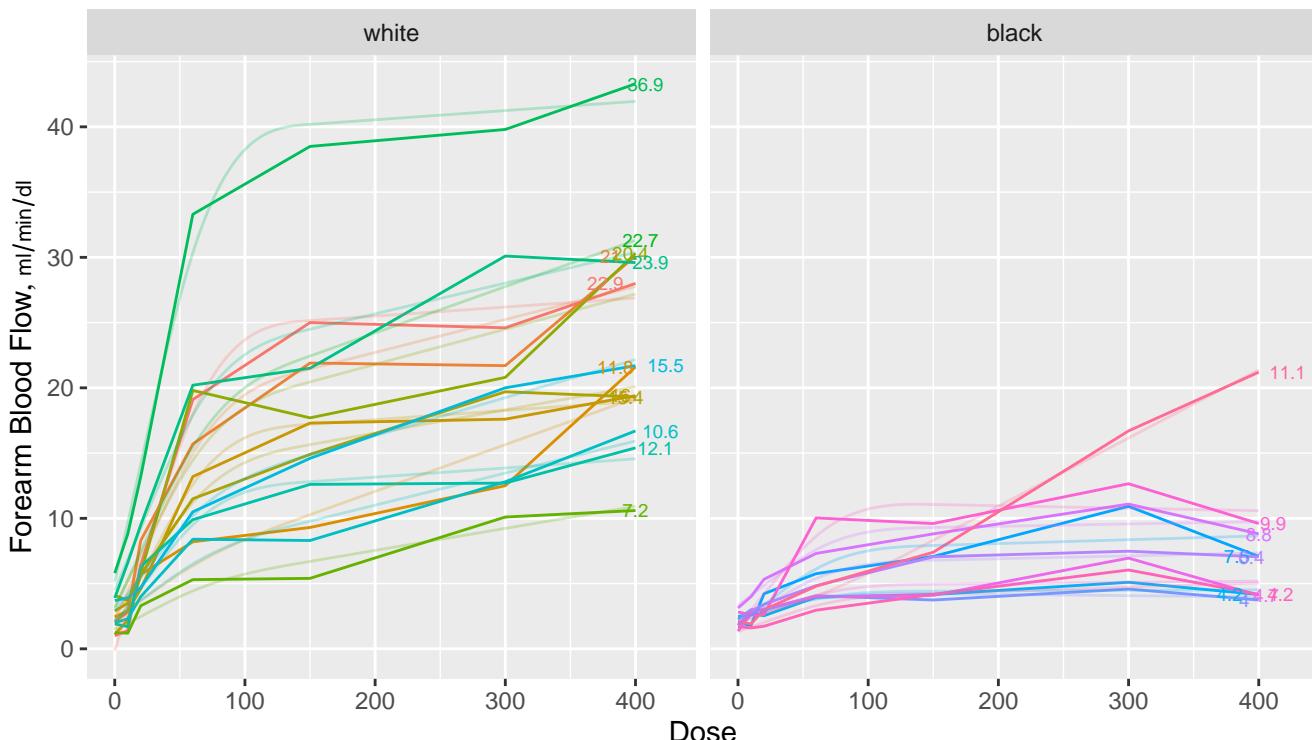


Figure 15.1: Spaghetti plots for isoproterenol data showing raw data stratified by race. Next to each curve is the area under the curve divided by 400 to estimate the mean response function. The area is computed analytically from a restricted cubic spline function fitted separately to each subject's dose-response curve. Shadowing the raw data are faint lines depicting spline fits for each subject

```

ggplot(a, aes(x=tarea, y=area, color=race)) + geom_point() +
  geom_abline(col=gray(.8)) +

```

```
xlab('Area by Trapezoidal Rule / 400') +
ylab('Area by Spline Fit / 400') # Fig. 15.2
```

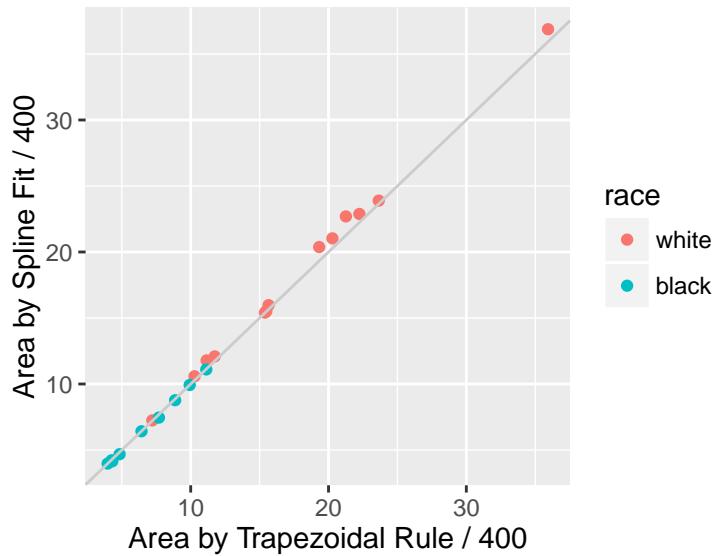


Figure 15.2: AUC by curve fitting and by trapezoidal rule

When a subject's dose (or time) span includes the minimum and maximum values over all subjects, one can use the trapezoidal rule to estimate the area under the response curve empirically. When interior points are missing, linear interpolation is used. The spline fits use nonlinear interpolation, which is slightly better, as is the spline function's assumption of continuity in the slope. Figure 15.2 compares the area under the curve (divided by 400 to estimate the mean response) estimated using the the two methods. Agreement is excellent. In this example, the principle advantage of the spline approach is that slope and shape parameters are estimated in the process, and these parameters may be tested separately for association with group (here, race). For example, one may test whether slopes differ across groups, and whether the means, curvatures, or inflection points differ. One could also compare AUC from a sub-interval of X .

```
ggplot(a, aes(x=race, y=area)) + # Fig. 15.3
  geom_boxplot(alpha=.5, width=.25) + geom_point() + coord_flip() +
  ylab(expression(paste('Mean Forearm Blood Flow, ', scriptstyle(ml/min/dl))))
```

L

K

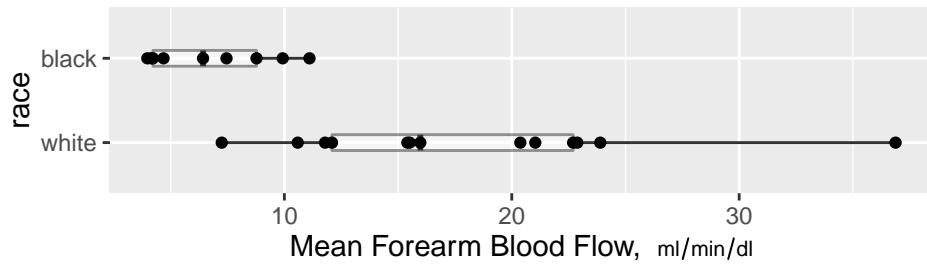


Figure 15.3: Mean blood flow computed from the areas under the spline curves, stratified by race, along with box plots

15.3.2

Nonparametric Test of Race Differences in AUC



A minimal-assumption approach to testing for differences in isoproterenol dose-response between races is to apply the Wilcoxon test to the normalized AUCs (mean response functions).

```
wilcox.test(area ~ race, data=a, conf.int=TRUE)
```

```
Wilcoxon rank sum test

data: area by race
W = 112, p-value = 7.639e-05
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 5.670113 16.348121
sample estimates:
difference in location
 11.25836
```

There is strong evidence that the mean response is greater for whites.

15.3.3

Nonparametric Test of General Curve Differences

N

O'Brien [76] proposed a method for using logistic regression to turn the Hotelling T^2 test on its side. The Hotelling test is the multivariate analog of the two-sample t -test, and can be used to test simultaneously such things as whether a treatment modifies either systolic or diastolic blood pressure. O'Brien's idea was to test whether systolic or diastolic blood pressure (or both) can predict which treatment was given. Here we use the idea to test for race differences in the shape of the dose-response curves. We do this by predicting race from a 3-predictor model—one containing the intercept, the next the coefficient of the linear dose effect and the third the coefficient of the nonlinear restricted cubic spline term (differences in cubes). These coefficients were estimated using ordinary least squares in separately predicting each subject's relationship between dose and forearm blood flow.

```
h ← d[, j=g(dose, fbf, what='coef'), by = list(id, race)]
h
```

	id	race	b0	b1	b2
1:	1	white	-0.1264763	0.32310327	-0.34253352

```

2: 2 white 2.1147707 0.22798336 -0.21959542
3: 3 white 2.3378721 0.06927717 -0.03625384
4: 4 white 1.4822502 0.19837740 -0.20727584
5: 5 white 2.5366751 0.15399740 -0.14756999
6: 6 white 3.2117187 0.19910942 -0.18650561
7: 7 white 1.4264366 0.05261565 -0.03871545
8: 8 white 3.0999999 0.21833887 -0.19813457
9: 9 white 5.1507764 0.45026617 -0.48013920
10: 10 white 4.4778127 0.23853904 -0.23289815
11: 11 white 1.9052885 0.13548226 -0.13917910
12: 12 white 2.1828176 0.07558431 -0.05524955
13: 13 white 2.9318982 0.12776900 -0.10679867
14: 14 black 2.3336099 0.02679742 -0.02856275
15: 15 black 1.8356227 0.07652884 -0.07972036
16: 16 black 2.5342537 0.02290717 -0.02585081
17: 17 black 2.0254606 0.06002835 -0.06261969
18: 18 black 3.3279080 0.07620477 -0.08062536
19: 19 black 1.9308650 0.03844018 -0.04060065
20: 20 black 1.7263259 0.12358392 -0.13595538
21: 21 black 1.3215502 0.03528716 -0.03480467
22: 22 black 2.0828281 0.03143768 0.02251155
      id   race      b0       b1       b2

```

```
f ← lrm(race ~ b0 + b1 + b2, data=h, x=TRUE, y=TRUE)
print(f, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = race ~ b0 + b1 + b2, data = h, x = TRUE, y = TRUE)
```

		Model Likelihood		Discrimination		Rank Discrim.	
		Ratio Test		Indexes		Indexes	
Obs	22	LR χ^2	19.73	R^2	0.798	C	0.957
white	13	d.f.	3	g	6.783	D_{xy}	0.915
black	9	$Pr(> \chi^2)$	0.0002	g_r	882.952	γ	0.915
$\max \left \frac{\partial \log L}{\partial \beta} \right 8 \times 10^{-7}$				g_p	0.471	τ_a	0.463
				Brier	0.080		

	Coef	S.E.	Wald Z	Pr(> Z)
Intercept	3.6618	4.1003	0.89	0.3718
b0	1.3875	2.5266	0.55	0.5829
b1	-165.3481	92.7864	-1.78	0.0747
b2	-101.1667	63.1549	-1.60	0.1092

The likelihood ratio $\chi^2_3 = 19.73$ has $P = 0.0002$ indicating strong evidence that the race variable has different averages, slopes, or shapes of the dose-response curves. The c -

index of 0.957 indicates nearly perfect ability to separate the races on the basis of three curve characteristics (although the sample size is small). We can use the bootstrap to get an overfitting-corrected index.

```
set.seed(2)
v ← validate(f, B=1000)
```

Divergence or singularity in 302 samples

```
latex(v, file='')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	<i>n</i>
D_{xy}	0.9145	0.9529	0.8476	0.1053	0.8092	698
R^2	0.7985	0.8987	0.6707	0.2280	0.5705	698
Intercept	0.0000	0.0000	0.1102	-0.1102	0.1102	698
Slope	1.0000	1.0000	0.4625	0.5375	0.4625	698
E_{\max}	0.0000	0.0000	0.1880	0.1880	0.1880	698
D	0.8513	1.0572	0.6556	0.4016	0.4498	698
U	-0.0909	-0.0909	2.4183	-2.5092	2.4183	698
Q	0.9422	1.1481	-1.7626	2.9107	-1.9685	698
B	0.0803	0.0345	0.0781	-0.0436	0.1239	698
g	6.7833	22.2084	4.5453	17.6631	-10.8798	698
g_p	0.4712	0.4629	0.4288	0.0341	0.4371	698

The overfitting-corrected c -index is $c = \frac{D_{xy}+1}{2} = 0.9$.

15.3.4

Model-Based Analysis: Generalized Least Squares



Generalized least squares (GLS) is the first generalization of ordinary least squares (multiple linear regression). It is described in detail in *Regression Modeling Strategies* Chapter 7 where a comprehensive case study is presented. The assumptions of GLS are

Q

- All the usual assumptions about the right-hand-side of the model related to transformations of X and interactions
- Residuals have a normal distribution
- Residuals have constant variance vs. \hat{Y} or any X (but the G in GLS also refers to allowing variances to change across X)

- The multivariate responses have a multivariate normal distribution conditional on X
- The correlation structure of the conditional multivariate distribution is correctly specified

With fully specified serial data models such as GLS, the fixed effects of time or dose are modeled just as any other predictor, with the only difference being that it is the norm to interact the time or dose effect with treatment or whatever X effect is of interest. This allows testing hypotheses such as

R

- Does the treatment effect change over time? (time \times treatment interaction)
- Is there a time at which there is a treatment effect? (time \times treatment interaction + treatment main effect combined into a chunk test)
- Does the treatment have an effect at time t ? (difference of treatments fixing time at t , not assuming difference is constant across different t)

In the majority of longitudinal clinical trials, the last hypothesis is the most important, taking t as the end of treatment point. This is because one is often interested in where patients ended up, not just whether the treatment provided temporary relief.

S

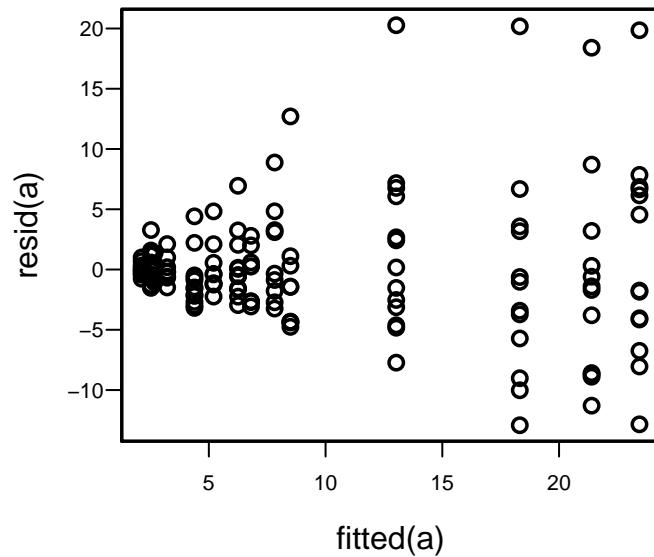
Now consider the isoproterenol dataset and fit a GLS model allowing for the same nonlinear spline effect of dose as was used above, and allowing the shapes of curves to be arbitrarily different by race. We impose a continuous time AR1 correlation structure on within-subject responses. This is the most commonly used correlation structure; it assumes that the correlation between two points is exponentially declining as the difference between two times or doses increases. We fit the GLS model and examine the equal variance assumption.

```
require(nlme)
```

```
dd <- datadist(d); options(datadist='dd')
a <- Glm(fbf ~ race * rcs(dose, c(20,60,150)), data=d,
          correlation=corCAR1(form = ~ dose | id))
plot(fitted(a), resid(a)) # Fig. 15.4
```

T

The variance of the residuals is clearly increasing with increasing dose. Try log transforming both `fbf` and dose. The log transformation requires a very arbitrary adjustment to dose to handle zeros.

Figure 15.4: Residual plot for generalized least squares fit on untransformed `fbf`

```
a <- Glm(log(fbf) ~ race * rcs(log(dose + 1), log(c(20,60,150)+1)), data=d,
           correlation=corCAR1(form = ~ dose | id))
print(anova(a), file='', table.env=FALSE)
```

	χ^2	d.f.	P
race (Factor+Higher Order Factors)	99.54	3	< 0.0001
All Interactions	19.30	2	0.0001
dose (Factor+Higher Order Factors)	312.10	4	< 0.0001
All Interactions	19.30	2	0.0001
Nonlinear (Factor+Higher Order Factors)	2.01	2	0.3667
race × dose (Factor+Higher Order Factors)	19.30	2	0.0001
Nonlinear	0.07	1	0.7969
Nonlinear Interaction : $f(A,B)$ vs. AB	0.07	1	0.7969
TOTAL NONLINEAR	2.01	2	0.3667
TOTAL NONLINEAR + INTERACTION	21.16	3	0.0001
TOTAL	391.48	5	< 0.0001

There is little evidence for a nonlinear dose effect on the log scale, implying that the underlying model is exponential on the original X and Y scales. This is consistent with Dupont [25]. Re-fit the model as linear in the logs. Before taking this as the final model, also fit the same model but using a correlation pattern based on time rather than dose. Assume equal time spacing during dose escalation.

```
a <- Glm(log(fbf) ~ race * log(dose + 1), data=d,
           correlation=corCAR1(form = ~ dose | id))
```

```
d$time <- match(d$dose, c(0, 10, 20, 60, 150, 300, 400)) - 1
b <- Glm(log(fbf) ~ race * log(dose + 1), data=d,
           correlation=corCAR1(form = ~ time | id))
AIC(a);AIC(b)
```

```
[1] 231.3731
```

```
[1] 161.3765
```

Lower AIC is better, so it is clear that time-based correlation structure is far superior to dose-based. We will used the second model for the remainder of the analysis. But first we check some of the model assumptions.

```
print(b, latex=TRUE)
```

Generalized Least Squares Fit by REML

```
Glm(model = log(fbf) ~ race * log(dose + 1), data = d, correlation = corCAR1
     id))
```

Obs	150	Log-restricted-likelihood	-74.69
Clusters	22	Model d.f.	3
<i>g</i>	0.755	σ	0.5023
		d.f.	146

	Coef	S.E.	t	Pr(> t)
Intercept	0.9851	0.1376	7.16	< 0.0001
race=black	-0.2182	0.2151	-1.01	0.3120
dose	0.3251	0.0286	11.38	< 0.0001
race=black * dose	-0.1421	0.0446	-3.19	0.0018

Correlation Structure: Continuous AR(1)

Formula: ~time | id

Parameter estimate(s):

Phi

0.6886846

```
latex(anova(b), file='', table.env=FALSE)
```

	χ^2	d.f.	P
race (Factor+Higher Order Factors)	32.45	2	< 0.0001
<i>All Interactions</i>	10.16	1	0.0014
dose (Factor+Higher Order Factors)	158.11	2	< 0.0001
<i>All Interactions</i>	10.16	1	0.0014
race \times dose (Factor+Higher Order Factors)	10.16	1	0.0014
TOTAL	180.17	3	< 0.0001

```
w <- data.frame(residual=resid(b), fitted=fitted(b))
p1 <- ggplot(w, aes(x=fitted, y=residual)) + geom_point()
p2 <- ggplot(w, aes(sample=residual)) + stat_qq()
gridExtra::grid.arrange(p1, p2, ncol=2) # Figure 15.5
```

V

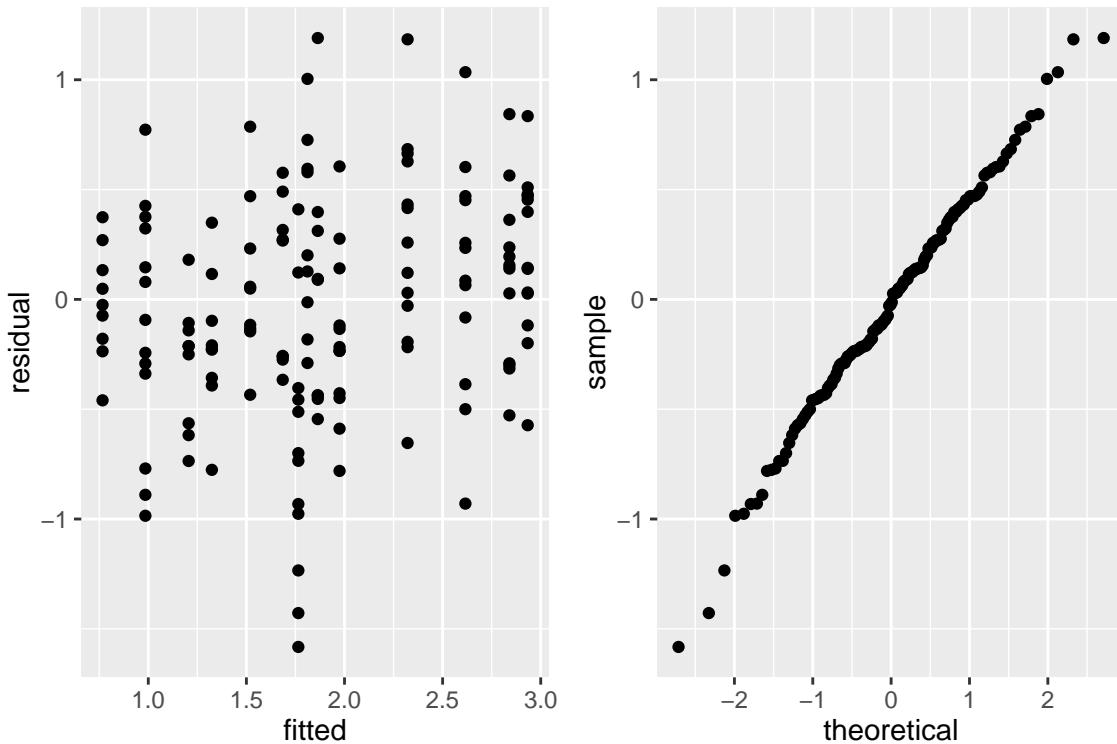


Figure 15.5: Checking assumptions of the GLS model that is linear after logging dose and blood flow. The graph on the right is a QQ-plot to check normality of the residuals from the model, where linearity implies normality.

The test for dose \times race interaction in the above ANOVA summary of Wald statistics shows strong evidence for difference in curve characteristics across races. This test agrees in magnitude with the less parametric approach using the logistic model above. But the logistic model also tests for an overall shift in distributions due to race, and the more efficient test for the combined race main effect and dose interaction effect from GLS is more significant with a Wald $\chi^2 = 32.45^c$. The estimate of the correlation between two log blood flows measured on the same subject one time unit apart is 0.69.

W

^cThe χ^2 for race with the dose-based correlation structure was a whopping 100 indicating that lack of fit of the correlation structure can have a significant effect on the rest of the GLS model.

The equal variance and normality assumptions appear to be well met as judged by Figure 15.5.

Now estimate the dose-response curves by race, with pointwise confidence intervals and simultaneous intervals that allow one to make statements about the entire curves^d. Anti-log the predicted values to get predictions on the original blood flow scale. Anti-logging predictions from a model that assumes a normal distribution on the logged values results in estimates of the median response.

```
dos ← seq(0, 400, length=150)
p ← Predict(b, dose=dos, race, fun=exp)
s ← Predict(b, dose=dos, race, fun=exp, conf.type='simultaneous')
ps ← rbind(Pointwise=p, Simultaneous=s)
ggplot(ps, ylab=expression(paste('Median Forearm Blood Flow, ', 
scriptstyle(ml/min/dl)))) # Fig. 15.6
```

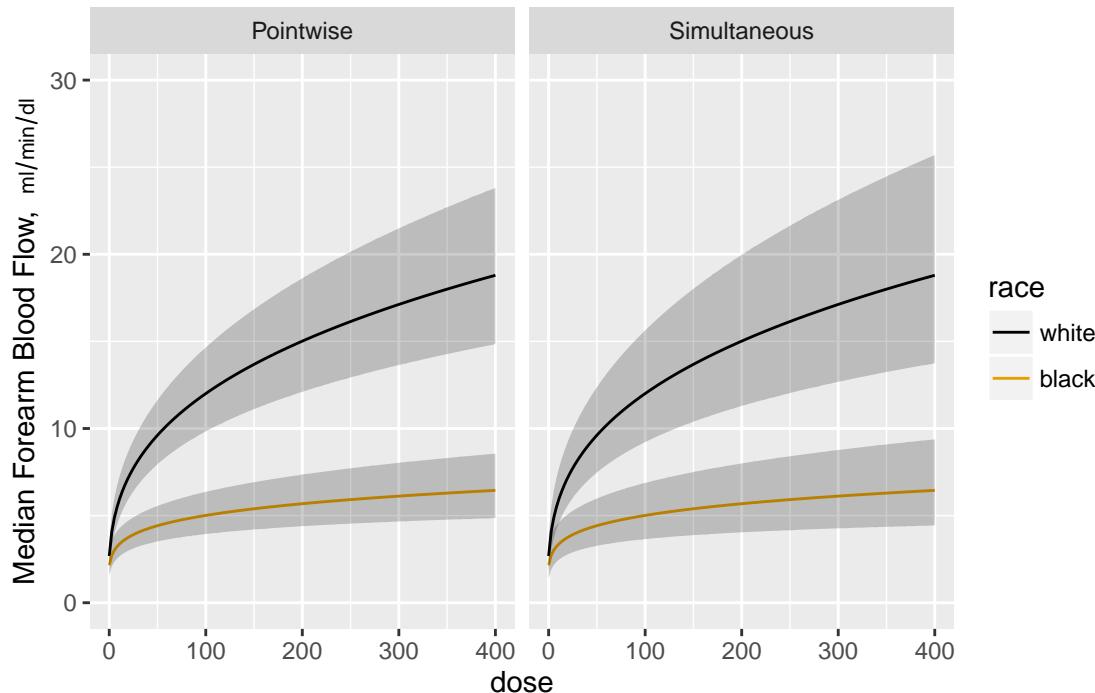

X


Figure 15.6: Pointwise and simultaneous confidence bands for median dose-response curves by race

Finally we estimate the white:black fold change (ratio of medians) as a function of dose with simultaneous confidence bands.

```
k ← contrast(b, list(dose=dos, race='white'),
list(dose=dos, race='black'), conf.type='simultaneous')
k ← as.data.frame(k[c('dose', 'Contrast', 'Lower', 'Upper')])
ggplot(k, aes(x=dose, y=exp(Contrast))) + geom_line() +
geom_ribbon(aes(ymin=exp(Lower), ymax=exp(Upper)), alpha=0.2, linetype=0,
show_guide=FALSE) +
geom_hline(yintercept=1, col='red', size=.2) +
```

^dSince the model is linear in log dose there are two parameters involving dose—the dose main effect and the race \times dose interaction effect. The simultaneous inference adjustment only needs to reflect simultaneity in two dimensions.

```
ylab('White:Black Ratio of Median FBF')      # Fig. 15.7
```

Y

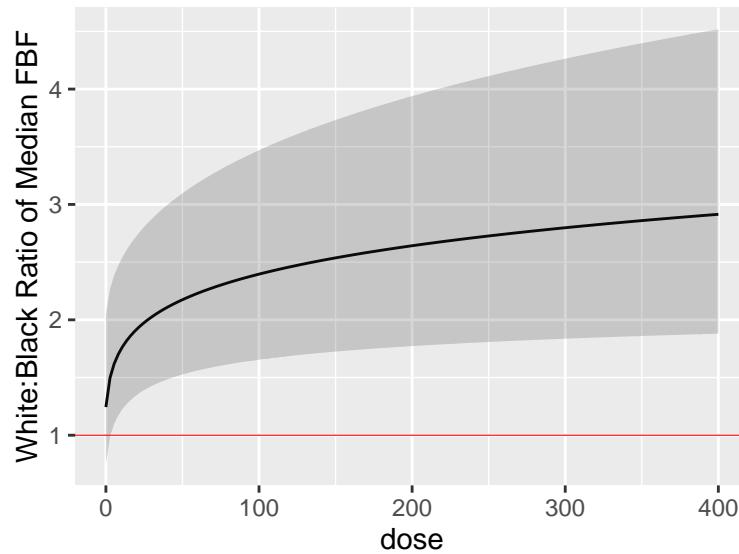


Figure 15.7: White:black fold change for median response as a function of dose, with simultaneous confidence band

By comparing the simultaneous confidence binds to the red horizontal line, one can draw the inference that the dose-response for blacks is everywhere different than that for whites when the dose exceeds zero.

Chapter 16

Analysis of Observer Variability and Measurement Agreement

16.1

Intra- and Inter-observer Disagreement

Before using a measurement instrument or diagnostic technique routinely, a researcher may wish to quantify the extent to which two determinations of the measurement, made by two different observers or measurement devices, disagree (inter-observer variability). She may also wish to quantify the repeatability of one observer in making the measurement at different times (intra-observer variability). To make these assessments, she has each observer make the measurement for each of a number of experimental units (e.g., subjects).

The measurements being analyzed may be continuous, ordinal, or binary (yes/no). Ordinal measurements must be coded such that distances between values reflects the relative importance of disagreement. For example, if a measurement has the values 1, 2, 3 for poor, fair, good, it is assumed that "good" is as different from "fair" as "fair" is from "poor". If this is not the case, a different coding should be used, such as coding 0 for "poor" if poor should be twice as far from "fair" as "fair" is from "good". Measurements that are yes/no or positive/negative should be coded as 1 or 0. The reason for this will be seen below.

There are many statistical methods for quantifying inter- and intra-observer variability. Correlation coefficients are frequently reported, but a perfect correlation can result

even when the measurements disagree by a factor of 10. Variance components analysis and intra-class correlation are often used, but these make many assumptions, do not handle missing data very well, and are difficult to interpret. Some analysts, in assessing inter-observer agreement when each observer makes several determinations, compute differences between the average determinations for each observer. This method clearly yields a biased measurement of inter-observer agreement because it cancels the intra-observer variability.

A general and descriptive method for assessing observer variability will now be presented. The methods uses a general type of statistic called the U statistic, invented by Hoeffding^{46a}. For definiteness, an analysis for 3 observers and 2 readings per observer will be shown. When designing such a study, the researcher should remember that the number of experimental units is usually the critical factor in determining the precision of estimates. There is not much to gain from having each observer make more than a few readings or from having 30 observers in the study (although if few observers are used, these are assumed to be “typical” observers).

The intra-observer disagreement for a single subject or unit is defined as the average of the intra-observer absolute measurement differences. In other words, intra-observer disagreement is the average absolute difference between any two measurements from the same observer. The inter-observer disagreement for one unit is defined as the average absolute difference between any two readings from different observers. Disagreement measures are computed separately for each unit and combined over units (by taking the mean or median for example) to get an overall summary measure. Units having more readings get more weight. When a reading is missing, that reading does not enter into any calculation and the denominator used in finding the mean disagreement is reduced by one.

Suppose that for one patient, observers A, B, and C make the following determinations on two separate occasions, all on the same patient:

A	B	C
5,7	8,5	6,7

For that patient, the mean intra-observer difference is $(|5 - 7| + |8 - 5| + |6 - 7|)/3 = \frac{2+3+1}{3} = 2$. The mean inter-observer difference is $(|5 - 8| + |5 - 5| + |5 - 6| + |5 - 7| + |7 - 8| + |7 - 5| + |7 - 6| + |7 - 7| + |8 - 6| + |8 - 7| + |5 - 6| + |5 - 7|)/12 = (3 + 0 + 1 + 2 + 1 + 2 + 1 + 0 + 2 + 1 + 1 + 2)/12 = \frac{16}{12} = 1.33$. If the first reading for observer A were unobtainable, the mean intra-observer difference for that patient

^aThe Wilcoxon test and the c -index are other examples of U statistics.

would be $(|8 - 5| + |6 - 7|)/2 = \frac{3+1}{2} = 2$ and the mean inter-observer difference would be $(|7 - 8| + |7 - 5| + |7 - 6| + |7 - 7| + |8 - 6| + |8 - 7| + |5 - 6| + |5 - 7|)/8 = (1 + 2 + 1 + 0 + 2 + 1 + 1 + 2)/8 = \frac{10}{8} = 1.25$.

The computations are carried out in like manner for each patient and summarized as follows:

Patient	Intra-observer Difference	Inter-observer Difference
1	2.00	1.33
2	1.00	3.50
3	1.50	2.66
.	.	.
.	.	.
<i>n</i>	.	.
Overall Average (or median)	1.77	2.23
Q_1	0.30	0.38
Q_3	2.15	2.84

Here is an example using R to compute mean inter- and intra-observer absolute differences for 4 subjects each assessed twice by each of 3 observers. The first subject consists of data above. The calculations are first done for the first subject alone, to check against computations above.

```
d <- expand.grid(rep=1:2, observer=c('A','B','C'), subject=1:4)
d$y <- c(5,7, 8,5, 6,7,
        7,6, 8,6, 9,7,
        7,5, 4,6, 10,11,
        7,6, 5,6, 9,8)
d
```

	rep	observer	subject	y
1	1	A	1	5
2	2	A	1	7
3	1	B	1	8
4	2	B	1	5
5	1	C	1	6
6	2	C	1	7
7	1	A	2	7
8	2	A	2	6
9	1	B	2	8
10	2	B	2	6
11	1	C	2	9
12	2	C	2	7
13	1	A	3	7
14	2	A	3	5
15	1	B	3	4

16	2	B	3	6
17	1	C	3	10
18	2	C	3	11
19	1	A	4	7
20	2	A	4	6
21	1	B	4	5
22	2	B	4	6
23	1	C	4	9
24	2	C	4	8

```
# Function to compute mean absolute discrepancies
mad <- function(y, obs, subj) {
  nintra <- ninter <- sumintra <- suminter <- 0
  n <- length(y)
  for(i in 1 : (n - 1)) {
    for(j in (i + 1) : n) {
      if(subj[i] == subj[j]) {
        dif <- abs(y[i] - y[j])
        if(! is.na(dif)) {
          if(obs[i] == obs[j]) {
            nintra <- nintra + 1
            sumintra <- sumintra + dif
          }
          else {
            ninter <- ninter + 1
            suminter <- suminter + dif
          }
        }
      }
    }
  }
  c(nintra=nintra, intra=sumintra / nintra,
    ninter=ninter, inter=suminter / ninter)
}

# Compute statistics for first subject
with(subset(d, subject == 1), mad(y, observer, subject))
```

nintra	intra	ninter	inter
3.000000	2.000000	12.000000	1.333333

```
# Compute for all subjects
with(d, mad(y, observer, subject))
```

nintra	intra	ninter	inter
12.000000	1.583333	48.000000	2.125000

Zhouwen Liu in the Vanderbilt Department of Biostatistics has developed much more general purpose software for this in R. Its web pages are <http://biostat.mc.vanderbilt.edu/AnalysisOfObserverVariability> and <https://github.com/harrelfe/rscripts>. The following example loads the source code and runs the above example. The R functions implement bootstrap nonparametric percentile confidence limits for mean absolute discrepancy measures.

```
require(Hmisc)
getRs('observerVariability.r', put='source')
```

YOU ARE RUNNING A DEMO VERSION 3_2

```
with(d, {
  intra ← intraVar(subject, observer, y)
  print(intra)
  summary(intra)
  set.seed(2)
  b=bootStrap(intra, by = 'subject', times=1000)
  # Get 0.95 CL for mean absolute intra-observer difference
  print(quantile(b, c(0.025, 0.975)))
  inter ← interVar(subject, observer, y)
  print(inter)
  summary(inter)
  b ← bootStrap(inter, by = 'subject', times=1000)
  # Get 0.95 CL for mean absolute inter-observer difference
  print(quantile(b, c(0.025, 0.975)))
})
```

Intra-Variability

Measures by Subjects and Raters

	subject	rater	variability	N
1	1	A	2	1
2	1	B	3	1
3	1	C	1	1
4	2	A	1	1
5	2	B	2	1
6	2	C	2	1
7	3	A	2	1
8	3	B	2	1
9	3	C	1	1
10	4	A	1	1
11	4	B	1	1
12	4	C	1	1

Measures by Subjects

	subject	variability	N
1	1	2.000000	3
2	2	1.666667	3
3	3	1.666667	3
4	4	1.000000	3

Measures by Raters

	subject	variability	N
1	A	1.50	4
2	B	2.00	4
3	C	1.25	4

Intra-variability summary:

Measures by Subjects and Raters
 Variability mean: 1.583333
 Variability median: 1.5
 Variability range: 1 3

```

Variability quantile:
0%: 1 25%: 1 50%: 1.5 75%: 2 100%: 3

Measures by Subjects
Variability mean: 1.583333
Variability median: 1.666667
Variability range: 1 2
Variability quantile:
0%: 1 25%: 1.5 50%: 1.666667 75%: 1.75 100%: 2

Measures by Raters
Variability mean: 1.583333
Variability median: 1.5
Variability range: 1.25 2
Variability quantile:
0%: 1.25 25%: 1.375 50%: 1.5 75%: 1.75 100%: 2
    2.5% 97.5%
1.166667 1.916667
Input object size:      3624 bytes;      9 variables      12 observations
Renamed variable       rater to rater1
Renamed variable       disagreement to variability
Dropped variables      diff, std, auxA1, auxA2
New object size:       1592 bytes;      5 variables      12 observations
Inter-Variability

Measures for All Pairs of Raters
  subject rater1 rater2 variability N
  1        1        1        2        1.5 4
  5        2        1        2        1.0 4
  9        3        1        2        1.5 4
 13        4        1        2        1.0 4
 17        1        1        3        1.0 4
 21        2        1        3        1.5 4
 25        3        1        3        4.5 4
 29        4        1        3        2.0 4
 33        1        2        3        1.5 4
 37        2        2        3        1.5 4
 41        3        2        3        5.5 4
 45        4        2        3        3.0 4

Measures by Rater Pairs
  rater1 rater2 variability N
  1        1        2        1.250 16
  2        1        3        2.250 16
  3        2        3        2.875 16

Measures by Subjects
  subject variability N
  1        1        1.333333 12
  2        2        1.333333 12
  3        3        3.833333 12
  4        4        2.000000 12
Inter-variability summary:
Measures for All Pairs of Raters
Variability mean: 2.125
Variability median: 1.5
Variability range: 1 5.5

```

```
Variability quantile:
0%: 1 25%: 1.375 50%: 1.5 75%: 2.25 100%: 5.5

Measures by Rater Pairs only
Variability mean: 2.125
Variability median: 2.25
Variability range: 1.25 2.875
Variability quantile:
0%: 1.25 25%: 1.75 50%: 2.25 75%: 2.5625 100%: 2.875

Measures by Subjects
Variability mean: 2.125
Variability median: 1.666667
Variability range: 1.333333 3.833333
Variability quantile:
0%: 1.333333 25%: 1.333333 50%: 1.666667 75%: 2.458333 100%: 3.833333
    2.5% 97.5%
1.333333 3.208333
```

```
# To load a demo file into an RStudio script editor window, type
# getRs('observerVariability_example.r')
```

From the above output, the 0.95 CL for the mean absolute intra-observer difference is [1.17, 1.92] and is [1.33, 3.21] for the inter-observer difference. The bootstrap confidence intervals use the cluster bootstrap to account for correlations of multiple readings from the same subject.

When the measurement of interest is a yes/no determination such as presence or absence of a disease these difference statistics are generalizations of the fraction of units in which there is exact agreement in the yes/no determination, when the absolute differences are summarized by averaging. To see this, consider the following data with only one observer:

Patient	Determinations	D_1, D_2	Agreement?	$ D_1 - D_2 $
1	YY	1 1	Y	0
2	YN	1 0	N	1
3	NY	0 1	N	1
4	NN	0 0	Y	0
5	NN	0 0	Y	0
6	YN	1 0	N	1

The average $|D_1 - D_2|$ is $\frac{3}{6} = 0.5$ which is equal to the proportion of cases in which the two readings disagree.

An advantage of this method of summarizing observer differences is that the investigator can judge what is an acceptable difference and he can relate this directly to the summary

disagreement statistic.

16.2

Comparison of Measurements with a Standard

When the true measurement is known for each unit (or the true diagnosis is known for each patient), similar calculations can be used to quantify the extent of errors in the measurements. For each unit, the average (over observers) difference from the true value is computed and these differences are summarized over the units. For example, if for unit #1 observer A measures 5 and 7, observer B measured 8 and 5, and the true value is 6, the average absolute error is $(|5 - 6| + |7 - 6| + |8 - 6| + |5 - 6|)/4 = \frac{1+1+2+1}{4} = \frac{5}{4} = 1.25$

16.3

Special Case: Assessing Agreement In Two Binary Variables

16.3.1

Measuring Agreement Between Two Observers

Suppose that each of n patients undergoes two diagnostic tests that can yield only the values positive and negative. The data can be summarized in the following frequency table.

		Test 2	
		+	-
Test 1	+	a	b
	-	c	d
		e	f
			g
			h
			n

An estimate of the probability that the two tests agree is $p_A = \frac{a+d}{n}$. An approximate 0.95 confidence interval for the true probability is derived from $p_A \pm 1.96\sqrt{p_A(1-p_A)/n}$ ^b. If the disease being tested is very rare or very common, the two tests will agree with high probability by chance alone. The κ statistic is one way to measure agreement that is corrected for chance agreement.

$$\kappa = \frac{p_A - p_C}{1 - p_C} \quad (16.1)$$

where p_C is the expected agreement proportion if the two observers are completely independent. The statistic can be simplified to

$$\kappa = \frac{2(ad - bc)}{gf + eh}. \quad (16.2)$$

If the two tests are in perfect agreement, $\kappa = 1$. If the two agree at the level expected by chance, $\kappa = 0$. If the level of agreement is less than one would obtain by chance alone, $\kappa < 0$.

A formal test of significance of the difference in the probabilities of for the two tests

^bA more accurate confidence interval can be obtained using Wilson's method as provided by the R Hmisc package binconf function.

is obtained using McNemar's test. The null hypothesis is that the probability of + for test 1 is equal to the probability of + for test 2, or equivalently that the probability of observing a +- is the same as that of observing -+. The normal deviate test statistic is given by

$$z = \frac{b - c}{\sqrt{b + c}}. \quad (16.3)$$

16.3.2

Measuring Agreement Between One Observer and a Standard

Suppose that each of n patients is studied with a diagnostic test and that the true diagnosis is determined, resulting in the following frequency table:

		Diagnosis	
		+	-
Test	+	a	b
	-	c	d
		e	f
			g
			h
			n

The following measures are frequently used to describe the agreement between the test and the true diagnosis. Here T^+ denotes a positive test, D^- denotes no disease, etc.

Quantity	Probability Being Estimated	Formula
Correct diagnosis probability	$\text{Prob}(T = D)$	$\frac{a+d}{n}$
Sensitivity	$\text{Prob}(T^+ D^+)$	$\frac{a}{e}$
Specificity	$\text{Prob}(T^- D^-)$	$\frac{d}{f}$
Accuracy of a positive test	$\text{Prob}(D^+ T_+)$	$\frac{a}{g}$
Accuracy of a negative test	$\text{Prob}(D^- T_-)$	$\frac{d}{h}$

The first and last two measures are usually preferred. Note that when the disease is very rare or very common, the correct diagnosis probability will be high by chance alone. Since the sensitivity and specificity are calculated conditional on the diagnosis, the prevalence of disease does not directly affect these measures. But sensitivity and specificity will vary with every patient characteristic related to the actual ignored severity of disease.

When estimating any of these quantities, Wilson confidence intervals are useful adjunct

statistics. A less accurate 0.95 confidence interval is obtained from $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$ where p is the proportion and n is its denominator.

16.4**Problems**

1. Three technicians, using different machines, make 3 readings each. For the data that follow, calculate estimates of inter- and intra-technician discrepancy.

Technician		
1	2	3
Reading	Reading	Reading
1 2 3	1 2 3	1 2 3
18 17 14	16 15 16	12 15 12
20 21 20	14 12	13
26 20 23	18 20	22 24
19 17	16	21 23
28 24	32 29	29 25

2. Forty-one patients each receive two tests yielding the frequency table shown below. Calculate a measure of agreement (or disagreement) along with an associated 0.95 confidence interval. Also calculate a chance-corrected measure of agreement. Test the null hypothesis that the the tests have the same probability of being positive and the same probability of being negative. In other words, test the hypothesis that the chance of observing $+-$ is the same as observing $-+$.

		Test 2	
		+	-
Test 1	+	29	8
	-	0	4

16.5

References

Landis JR, Koch GG: A review of statistical methods in the analysis of data arising from observer reliability studies (Part II), *Statistica Neerlandica* **29**:151-619 1975.

Landis JR, Koch GG: An application of hierarchical κ -type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33**:363-74, 1977.

Chapter 17

Modeling for Observational Treatment Comparisons

17.1

Propensity Score

- In observational studies comparing treatments, need to adjust for nonrandom treatment selection
- Number of confounding variables can be quite large
- May be too large to adjust for them using multiple regression, due to overfitting (may have more potential confounders than outcome events)
- Assume that all factors related to treatment selection that are prognostic are collected
- Use them in a flexible regression model to predict treatment actually received (e.g., logistic model allowing nonlinear effects)
- **Propensity score (PS)** = estimated probability of getting treatment B vs. treatment A
- Use of the PS allows one to aggressively adjust for confounders by simulating a randomized trial if all confounders are in the PS model

- Doing an adjusted analysis where the adjustment variable is the PS simultaneously adjusts for all the variables in the score *insofar* as confounding is concerned (but **not with regard to outcome heterogeneity**)
- If after adjusting for the score there were a residual imbalance for one of the variables, that would imply that the variable was not correctly modeled in the PS
- E.g.: after holding PS constant there are more subjects above age 70 in treatment B; means that age $>$ 70 is still predictive of treatment received after adjusting for PS, or age $>$ 70 was not modeled correctly.
- An additive (in the logit) model where all continuous baseline variables are splined will result in adequate adjustment in the majority of cases—certainly better than categorization. Lack of fit will then come only from omitted interaction effects. E.g.: if older males are much more likely to receive treatment B than treatment A than what would be expected from the effects of age and sex alone, adjustment for the additive propensity would not adequately balance for age and sex.

17.2

Assessing Treatment Effect

- Eliminate patients in intervals of PS where there is no overlap between A and B, or include an interaction between treatment and a baseline characteristic^a
- Many researchers stratify the PS into quintiles, get treatment differences within the quintiles, and average these to get adjustment treatment effects
- Often results in imbalances in outer quintiles due to skewed distributions of PS there
- Can do a matched pairs analysis but depends on matching tolerance and many patients will be discarded when their case has already been matched
- Usually better to adjust for PS in a regression model
- Model: $Y = \text{treat} + \log \frac{PS}{1-PS} + \text{nonlinear functions of } \log \frac{PS}{1-PS} + \text{important prognostic variables}$
- Prognostic variables need to be in outcome (Y) model even though they are also in the PS, to account for subject outcome heterogeneity (susceptibility bias)
- If outcome is binary and can afford to ignore prognostic variables, use nonparametric regression to relate PS to outcome separately in actual treatment A vs. B groups
- Plotting these two curves with PS on x -axis and looking at vertical distances between curves is an excellent way to adjust for PS continuously without assuming a model

^aTo quote Gelman and Hill Section 10.3³², "Ultimately, one good solution may be a multilevel model that includes treatment interactions so that inferences explicitly recognize the decreased precision that can be obtained outside the region of overlap." For example, if one included an interaction between age and treatment and there were no patients greater than 70 years old receiving treatment B, the B:A difference for age greater than 70 would have an extremely wide confidence interval as it depends on extrapolation. So the estimates that are based on extrapolation are not misleading; they are just not informative.

17.2.1

Problems with Propensity Score Matching

- The choice of the matching algorithm is not principle-based so is mainly arbitrary. Most matching algorithms are dependent on the order of observations in the dataset. Arbitrariness of matching algorithms creates a type of non-reproducibility.
- Non-matched observations are discarded, resulting in a loss of precision and power.
- Matching not only discards hard-to-match observations (thus helping the analyst correctly concentrate on the propensity overlap region) but also discards many “good” matches in the overlap region.
- Matching does not do effective interpolation on the interior of the overlap region.
- The choice of the main analysis when matching is used is not well worked out in the statistics literature. Most analysts just ignore the matching during the outcome analysis.
- Even with matching one must use covariate adjustment for strong prognostic factors to get the right treatment effects, due to non-collapsibility of odds and hazards ratios.
- Matching hides interactions with treatment and covariates.

Most users of propensity score matching do not even entertain the notion that the treatment effect may interact with propensity to treat, much less entertain the thought of individual patient characteristics interacting with treatment.

17.3

Recommended Statistical Analysis Plan

1. Be very liberal in selecting a large list of potential confounder variables that are measured pre-treatment.
2. If the number of potential confounders is not large in comparison with the effective sample size, use direct covariate adjustment instead of propensity score adjustment. For example, if the outcome is binary and you have more than 5 events per covariate, full covariate adjustment probably works OK.
3. Model the probability of receiving treatment using a flexible statistical model that makes minimal assumptions (e.g., rich additive model that assumes smooth predictor effects). If there are more than two treatments, you will need as many propensity scores as there are treatments, less one, and all of the logic propensity scores will need to be adjusted for in what follows.
4. Examine the distribution of estimated propensity score separately for the treatment groups.
5. If there is a non-overlap region of the two distributions, and you don't want to use a more conservative interaction analysis (see below), exclude those subjects from the analysis. Recursive partitioning can be used to predict membership in the non-overlap region from baseline characteristics so that the research findings with regard to applicability/generalizability can be better understood.
6. Use covariate adjustment for propensity score for subjects in the overlap region. Expand logit propensity using a restricted cubic spline so as to not assume linearity in the logit in relating propensity to outcome. Also include pre-specified important prognostic factors in the model to account for the majority of outcome heterogeneity. It is not a problem that these prognostic variables are also in the propensity score.
7. As a secondary analysis use a chunk test to assess whether there is an interaction with logit propensity to treat and actual treatment. For example, one may find that physicians are correctly judging that one subset of patients should usually be treated a certain way.
8. Instead of removing subjects outside the overlap region, you could allow propensity or individual predictors to interact with treatment. Treatment effect estimates in the presence of interactions are self-penalizing for not having sufficient overlap.

Suppose for example that age were the only adjustment covariate and a propensity score was not needed. Suppose that for those with age less than 70 there were sufficiently many subjects from either treatment for every interval of age but that when age exceeded 70 there were only 5 subjects on treatment B. Including an age \times treatment interaction in the model and obtaining the estimated outcome difference for treatment A vs. treatment B as a function of age will have a confidence band with minimum width at the mean age, and above age 70 the confidence band will be very wide. This is to be expected and is an honest way to report what we know about the treatment effect adjusted for age. If there were no age \times treatment interaction, omitting the interaction term would yield a proper model with a relatively narrow confidence interval, and if the shape of the age relationship were correctly specified the treatment effect estimate would be valid. So one can say that not having comparable subjects on both treatments for some intervals of covariates means that either (1) inference should be restricted to the overlap region, or (2) the inference is based on model assumptions.

Using a full regression analysis allows interactions to be explored, as briefly described above. Suppose that one uses a restricted cubic spline in the logit propensity to adjust for confounding, and all these spline terms are multiplied by the indicator variable for getting a certain treatment. One can make a plot with predicted outcome on the y -axis and PS on the x -axis, with one curve per treatment. This allows inspection of parallelism (which can easily be formally tested with the chunk test) and whether there is a very high or very low PS region where treatment effects are different from the average effect. For example, if physicians have a very high probability of always selecting a certain treatment for patients that actually get the most benefit from the treatment, this will be apparent from the plot.

17.4

Reasons for Failure of Propensity Analysis

Propensity analysis may not sufficiently adjust for confounding in nonrandomized studies when

- prognostic factors that are confounders are not measured and are not highly correlated with factors that are measured
- the propensity modeling was too parsimonious (e.g., if the researchers excluded baseline variables just because they were insignificant)
- the propensity model assumed linearity of effects when some were really nonlinear (this would cause an imbalance in something other than the mean to not be handled)
- the propensity model should have had important interaction terms that were not included (e.g., if there is only an age imbalance in males)
- the researchers attempted to extrapolate beyond ranges of overlap in propensity scores in the two groups (this happens with covariate adjustment sometimes, but can happen with quantile stratification if outer quantiles are very imbalanced)

17.5

Sensitivity Analysis

- For n patients in the analysis, generate n random values of a hypothetical unmeasured confounder U
- Constrain U so that the effect of U on the response Y is given by an adjusted odds ratio of OR_Y and so that U 's distribution is unbalanced in group A vs. B to the tune of an odds ratio of OR_{treat} .
- Solve for how large OR_Y and OR_{treat} must be before the adjusted treatment effect reverses sign or changes in statistical significance
- The larger are OR_Y and OR_{treat} the less plausible it is that such an unmeasured confounder exists

See the R `rms` package `sensuc` function.

17.6

Reasons To Not Use Propensity Analysis

Chen et al.¹⁶ demonstrated advantages of using a unified regression model to adjust for “too many” predictors by using penalized maximum likelihood estimation, where the exposure variable coefficients are not penalized but all the adjustment variable coefficients have a quadratic (ridge) penalty.

17.7

Further Reading

Gelman has a nice chapter on causal inference and matching from Gelman and Hill³².

Chapter 18

Information Loss

... wherever nature draws unclear boundaries, humans are happy to curate

Alice Dreger, *Galileo's Middle Finger*

This material is from "Information Allergy" by FE Harrell, presented as the Vanderbilt Discovery Lecture 2007-09-13 and presented as invited talks at Erasmus University, Rotterdam, The Netherlands, University of Glasgow (Mitchell Lecture), Ohio State University, Medical College of Wisconsin, Moffitt Cancer Center, U. Pennsylvania, Washington U., NIEHS, Duke, Harvard, NYU, Michigan, Abbott Labs, Becton Dickinson, NIAID, Albert Einstein, Mayo Clinic, U. Washington, MBSW, U. Miami, Novartis. Material is added from "How to Do Bad Biomarker Research" by FE Harrell, presented at the NIH NIDDK Conference *Towards Building Better Biomarkers; Statistical Methodology*, 2014-12-02.

18.1

Information & Decision Making

What is **information**?

- Messages used as the basis for decision-making
- Result of processing, manipulating and organizing data in a way that adds to the receiver's knowledge
- Meaning, knowledge, instruction, communication, representation, and mental stimulus^a

Information resolves uncertainty.

Some types of information may be quantified in bits. A binary variable is represented by 0/1 in base 2, and it has 1 bit of information. This is the minimum amount of information other than no information. Systolic blood pressure measured accurately to the nearest 4mmHg has 6 binary digits—bits—of information ($\log_2 \frac{256}{4} = 6$). Dichotomizing blood pressure reduces its information content to 1 bit, resulting in enormous loss of precision and power.

Value of information: Judged by the variety of outcomes to which it leads.

Optimum decision making requires the maximum and most current information the decision maker is capable of handling

Some important decisions in biomedical and epidemiologic research and clinical practice:

- Pathways, mechanisms of action
- Best way to use gene and protein expressions to diagnose or treat
- Which biomarkers are most predictive and how should they be summarized?
- What is the best way to diagnose a disease or form a prognosis?

^apbs.org/weta, wikipedia.org/wiki/Information

- Is a risk factor causative or merely a reflection of confounding?
- How should patient outcomes be measured?
- Is a drug effective for an outcome?
- Who should get a drug?

18.1.1

Information Allergy

Failing to obtain key information needed to make a sound decision

- Not collecting important baseline data on subjects

Ignoring Available Information

- Touting the value of a new biomarker that provides less information than basic clinical data
- Ignoring confounders (alternate explanations)
- Ignoring subject heterogeneity
- Categorizing continuous variables or subject responses
- Categorizing predictions as “right” or “wrong”
- Letting fear of probabilities and costs/utilities lead an author to make decisions for individual patients

18.2

Ignoring Readily Measured Variables

Prognostic markers in acute myocardial infarction

c-index: concordance probability \equiv receiver operating characteristic curve or ROC area
 Measure of ability to discriminate death within 30d

Markers	<i>c</i> -index
CK-MB	0.63
Troponin T	0.69
Troponin T > 0.1	0.64
CK-MB + Troponin T	0.69
CK-MB + Troponin T + ECG	0.73
Age + sex	0.80
All	0.83

Ohman et al. [77]

Though not discussed in the paper, age and sex easily trump troponin T. One can also see from the *c*-indexes that the common dichotomization of troponin results in an immediate loss of information.

Inadequate adjustment for confounders: Greenland [37]

- Case-control study of diet, food constituents, breast cancer
- 140 cases, 222 controls
- 35 food constituent intakes and 5 confounders
- Food intakes are correlated
- Traditional stepwise analysis not adjusting simultaneously for all foods consumed
 \rightarrow 11 foods had $P < 0.05$
- Full model with all 35 foods competing \rightarrow 2 had $P < 0.05$

- Rigorous simultaneous analysis (hierarchical random slopes model) penalizing estimates for the number of associations examined → no foods associated with breast cancer

Ignoring subject variability in randomized experiments

- Randomization tends to balance measured and unmeasured subject characteristics across treatment groups
- Subjects vary widely within a treatment group
- Subject heterogeneity usually ignored
- False belief that balance from randomization makes this irrelevant
- Alternative: analysis of covariance
- If any of the baseline variables are predictive of the outcome, there is a gain in power for every type of outcome (binary, time-to-event, continuous, ordinal)
- Example for a binary outcome in Section [13.2.2](#)

18.3

Categorization: Partial Use of Information

Patient: What was my systolic BP this time?

MD: It was > 120

Patient: How is my diabetes doing?

MD: Your $\text{Hb}_{\text{A}1c}$ was > 6.5

Patient: What about the prostate screen?

MD: If you have average prostate cancer, the chance that $\text{PSA} > 5$ in this report is 0.6

Problem: Improper conditioning ($X > c$ instead of $X = x$) \rightarrow information loss; reversing time flow

Sensitivity: $P(\text{observed } X > c \text{ given unobserved } Y = y)$

18.3.1

Categorizing Continuous Predictors



- Many physicians attempt to find cutpoints in continuous predictor variables
- Mathematically such cutpoints cannot exist unless relationship with outcome is discontinuous
- Even if the cutpoint existed, it **must** vary with other patient characteristics, as optimal decisions are based on risk
- A simple 2-predictor example related to diagnosis of pneumonia will suffice
- It is **never** appropriate to dichotomize an input variable other than time. Dichotomization, if it must be done, should **only** be done on \hat{Y} . In other words, dichotomization is done as late as possible in decision making. When more than one continuous predictor variable is relevant to outcomes, the example below shows that it is mathematically incorrect to do a one-time dichotomization of a predictor.

As an analogy, suppose that one is using body mass index (BMI) by itself to make a decision. One would never categorize height and categorize weight to make the decision based on BMI. One could categorize BMI, if no other outcome predictors existed for the problem.

```
require(rms)
```

```
getHdata(ari)
r <- ari[ari$age >= 42, Cs(age, rr, pneu, coh, s2)]
abn.xray <- r$s2==0
r$coh <- factor(r$coh, 0:1, c('no cough', 'cough'))
f <- lrm(abn.xray ~ rcs(rr,4)*coh, data=r)
anova(f)
```

	Wald Statistics	Response: abn.xray		
Factor		Chi-Square	d.f.	P
rr (Factor+Higher Order Factors)	37.45	6	<.0001	
All Interactions	0.35	3	0.9507	
Nonlinear (Factor+Higher Order Factors)	3.27	4	0.5144	
coh (Factor+Higher Order Factors)	28.91	4	<.0001	
All Interactions	0.35	3	0.9507	
rr * coh (Factor+Higher Order Factors)	0.35	3	0.9507	
Nonlinear	0.31	2	0.8549	
Nonlinear Interaction : f(A,B) vs. AB	0.31	2	0.8549	
TOTAL NONLINEAR	3.27	4	0.5144	
TOTAL NONLINEAR + INTERACTION	3.37	5	0.6431	
TOTAL	66.06	7	<.0001	

```
dd <- datadist(r); options(datadist='dd')
p <- Predict(f, rr, coh, fun=plogis, conf.int=FALSE)
ggplot(p, rdata=r,      # Fig. 18.1
       ylab='Probability of Pneumonia',
       xlab='Adjusted Respiratory Rate/min.',
       ylim=c(0,.7), legend.label='')
```

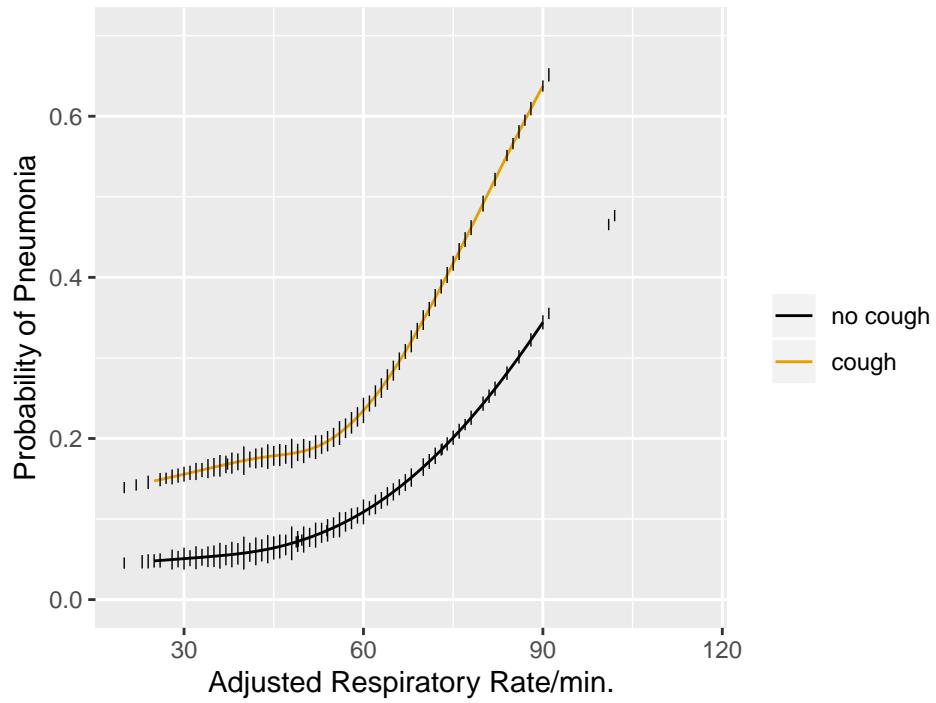


Figure 18.1: Estimated risk of pneumonia with respect to two predictors in WHO ARI study from Harrell et al. [42]. Tick marks show data density of respiratory rate stratified by cough. Any cutpoint for the rate **must** depend on cough to be consistent with optimum decision making, which must be risk-based.

18.3.2

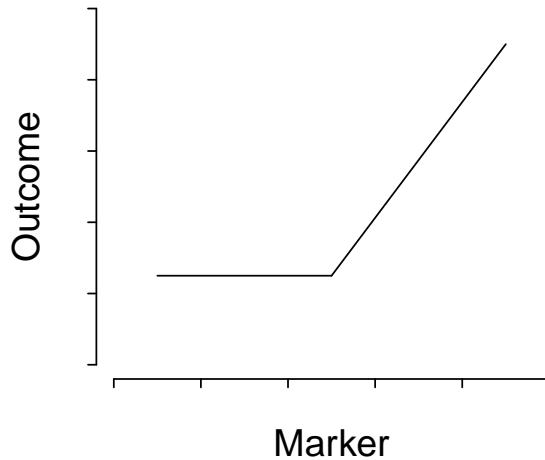
What Kinds of True Thresholds Exist?

Natura non facit saltus
(Nature does not make jumps)

Gottfried Wilhelm Leibniz

Can Occur in Biology

Not Handled by Dichotomization



Cannot Occur Unless $X = \text{time}$

Assumed in Much of Biomarker Research

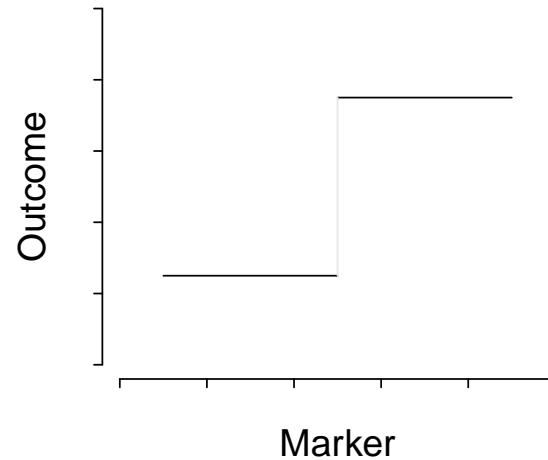


Figure 18.2: Two kinds of thresholds. The pattern on the left represents a discontinuity in the first derivative (slope) of the function relating a marker to outcome. On the right there is a lowest-order discontinuity.

What Do Cutpoints Really Assume?

Cutpoints assume discontinuous relationships of the type in the right plot of Figure 18.2, and they assume that the true cutpoint is known. Beyond the molecular level, such patterns do not exist unless $X = \text{time}$ and the discontinuity is caused by an event. Cutpoints assume homogeneity of outcome on either side of the cutpoint.

18.3.3

Cutpoints are Disasters

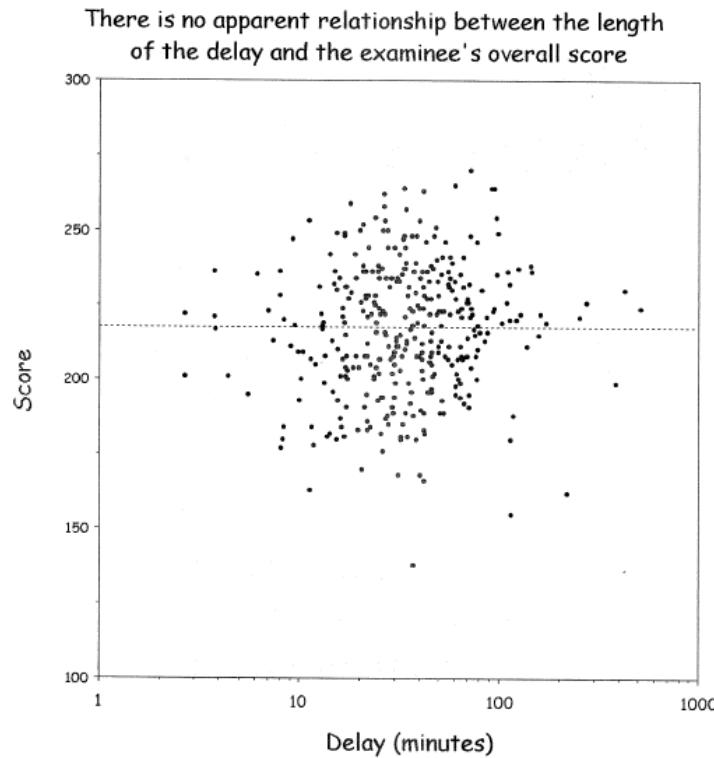
- Prognostic relevance of S-phase fraction in breast cancer: 19 different cutpoints used in literature
- Cathepsin-D content and disease-free survival in node-negative breast cancer: 12 studies, 12 cutpoints

- ASCO guidelines: neither cathepsin-D nor S-phrase fraction recommended as prognostic markers Holländer, Sauerbrei, and Schumacher [47]

Cutpoints may be found that result in both increasing and decreasing relationships with **any** dataset with zero correlation

Range of Delay	Mean Score	Range of Delay	Mean Score
0-11	210	0-3.8	220
11-20	215	3.8-8	219
21-30	217	8-113	217
31-40	218	113-170	215
41-	220	170-	210

Wainer [110]; See “Dichotomania” Senn [93] and Royston, Altman, and Sauerbrei [85]

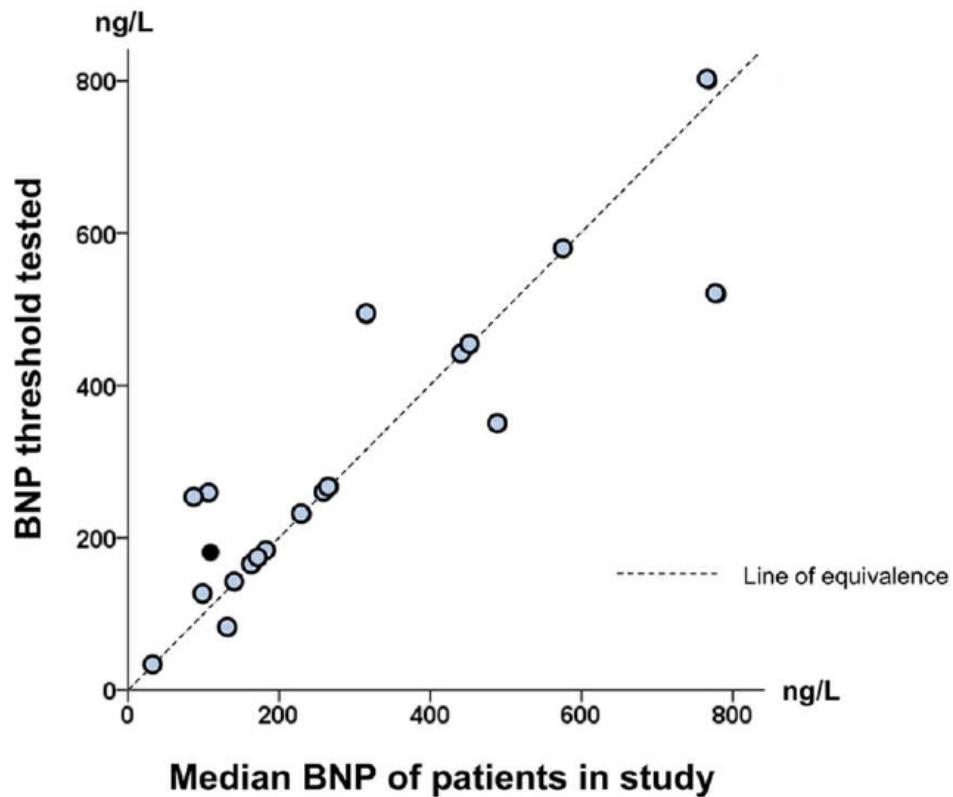


Wainer [110]

In fact, virtually all published cutpoints are analysis artifacts caused by finding a threshold that minimizes P -values when comparing outcomes of subjects below with those above the “threshold”. Two-sample statistical tests suffer the least loss of power when cutting at the median because this balances the sample sizes. That this method has nothing to do with biology can be readily seen by adding observations on either

tail of the marker, resulting in a shift of the median toward that tail even though the relationship between the continuous marker and the outcome remains unchanged.

BNP studies testing a single prognostic threshold, finding it to be either significant (○) or non significant (●)

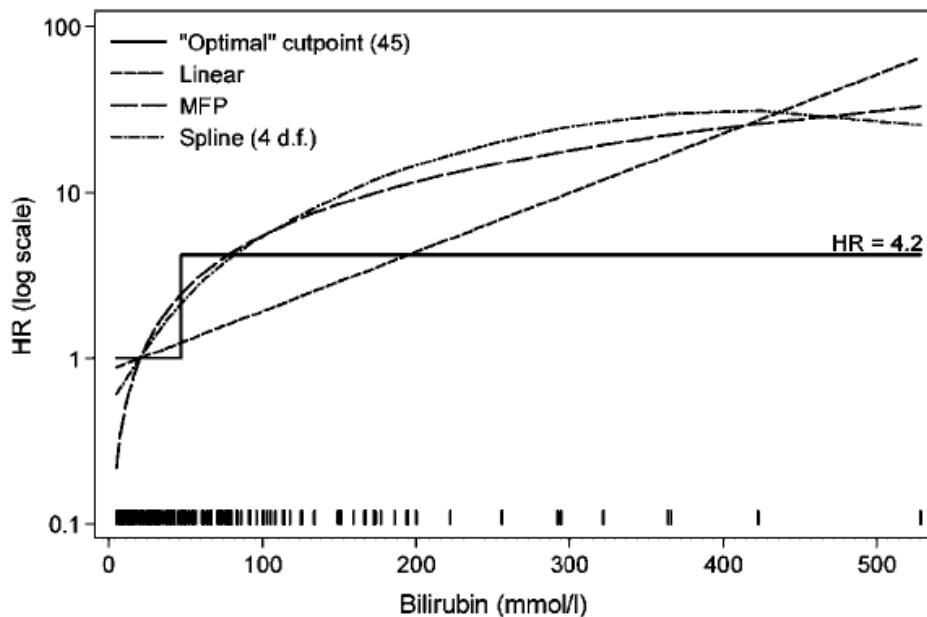


In “positive” studies: threshold 132–800 ng/L, correlation with study median $r = 0.86$ (Giannoni et al. [33])

Lack of Meaning of Effects Based on Cutpoints

- Researchers often use cutpoints to estimate the high:low effects of risk factors (e.g., BMI vs. asthma)
 - Results in inaccurate predictions, residual confounding, impossible to interpret
 - high:low represents unknown mixtures of highs and lows

- Effects (e.g., odds ratios) will vary with population
- If the true effect is monotonic, adding subjects in the low range or high range or both will increase odds ratios (and all other effect measures) arbitrarily



Royston, Altman, and Sauerbrei [85], Naggara et al. [72], Giannoni et al. [33]

Does a physician ask the nurse “Is this patient’s bilirubin > 45 ” or does she ask “What is this patient’s bilirubin level?”. Imagine how a decision support system would trigger vastly different decisions just because bilirubin was 46 instead of 44.

As an example of how a hazard ratio for a dichotomized continuous predictor is an arbitrary function of the entire distribution of the predictor within the two categories, consider a Cox model analysis of simulated age where the true effect of age is linear. First compute the $\geq 50 : < 50$ hazard ratio in all subjects, then in just the subjects having age < 60 , then in those with age < 55 . Then repeat including all older subjects but excluding subjects with age ≤ 40 . Finally, compute the hazard ratio when only those age 40 to 60 are included. Simulated times to events have an exponential distribution, and proportional hazards holds.

```
set.seed(1)
n <- 1000
age <- rnorm(n, mean=50, sd=12)
describe(age)
```

age	n	missing	distinct	Info	Mean	Gmd	.05	.10
	1000	0	1000	1	49.86	14.01	29.28	33.93
	.25	.50	.75	.90	.95			

```

41.63      49.58      58.26      65.89      70.93

lowest : 13.90342 14.03661 14.72272 15.33295 18.84666
highest: 79.97194 81.79000 82.10889 86.66891 95.72332

```

```

cens ← 15 * runif(n)
h ← 0.02 * exp(0.04 * (age - 50))
dt ← -log(runif(n))/h
e ← ifelse(dt ≤ cens, 1, 0)
dt ← pmin(dt, cens)
S ← Surv(dt, e)
coef(cph(S ~ age))    # close to true value of 0.04 used in simulation

```

```

age
0.04027519

```

```
exp(coef(cph(S ~ age ≥ 50)))    # ≥ 50 : < 50 hazard ratio estimate
```

```

age
2.148554

```

```
exp(coef(cph(S ~ age ≥ 50, subset=age < 60)))
```

```

age
1.645141

```

```
exp(coef(cph(S ~ age ≥ 50, subset=age < 55)))
```

```

age
1.461928

```

```
exp(coef(cph(S ~ age ≥ 50, subset=age > 40)))
```

```

age
1.760201

```

```
exp(coef(cph(S ~ age ≥ 50, subset=age > 40 & age < 60)))
```

```

age
1.354001

```

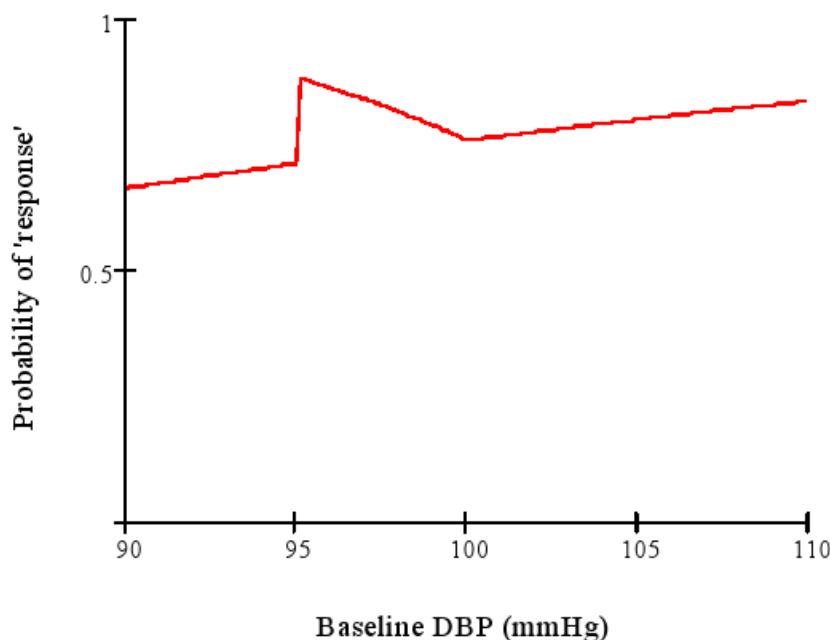
See [this](#) for excellent graphical examples of the harm of categorizing predictors, especially when using quantile groups.

18.3.4

Categorizing Outcomes

- Arbitrary, low power, can be difficult to interpret

- Example: “The treatment is called successful if either the patient has gone down from a baseline diastolic blood pressure of ≥ 95 mmHg to ≤ 90 mmHg or has achieved a 10% reduction in blood pressure from baseline.”
- Senn derived the response probability function for this discontinuous concocted endpoint



Senn [93] after Goetghebeur [1998]

Is a mean difference of 5.4mmHg more difficult to interpret than A:17% vs. B:22% hit clinical target?

“Responder” analysis in clinical trials results in huge information loss and arbitrariness.
Some issue:

- Responder analyses use cutpoints on continuous or ordinal variables and cite earlier data supporting their choice of cutpoints. No example has been produced where the earlier data actually support the cutpoint.
- Many responder analyses are based on change scores when they should be based solely on the follow-up outcome variable, adjusted for baseline as a covariate.
- The cutpoints are always arbitrary.
- There is a huge power loss (see Section 18.3.4).

- The responder probability is often a function of variables that one does not want it to be a function of (see graph above).

Fedorov, Mannino, and Zhang [28] is one of the best papers quantifying the information and power loss from categorizing continuous outcomes. One of their examples is that a clinical trial of 100 subjects with continuous Y is statistically equivalent to a trial of 158 dichotomized observations, assuming that the dichotomization is at the **optimum** point (the population median). They show that it is very easy for dichotomization of Y to raise the needed sample size by a factor of 5.

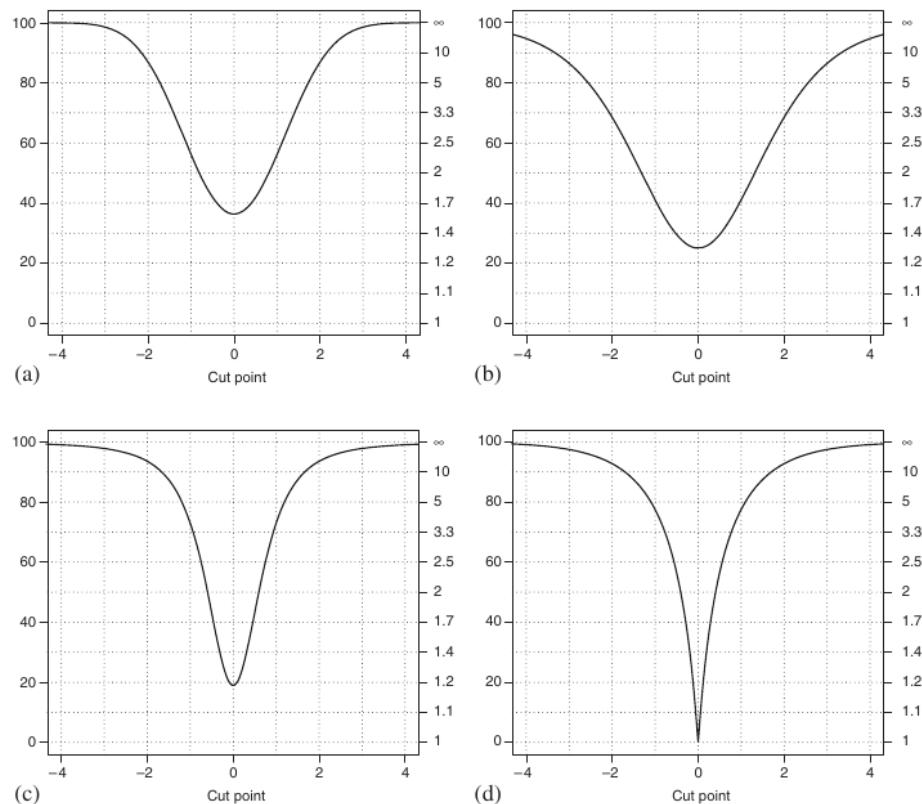


Figure 1. Percentage of information lost when dichotomizing, based on various cut points. The plots correspond to normal (a), logistic (b), Cauchy (c), and double exponential (d) distributions. The right vertical axes correspond to the factor of sample size increase needed to mitigate the loss.

Fedorov, Mannino, and Zhang [28]

18.3.5

Classification vs. Probabilistic Thinking



Number needed to treat. The only way, we are told, that physicians can understand probabilities: odds being a difficult concept only comprehensible to statisticians, bookies, punters and readers of the sports pages of popular newspapers.

Senn [92]

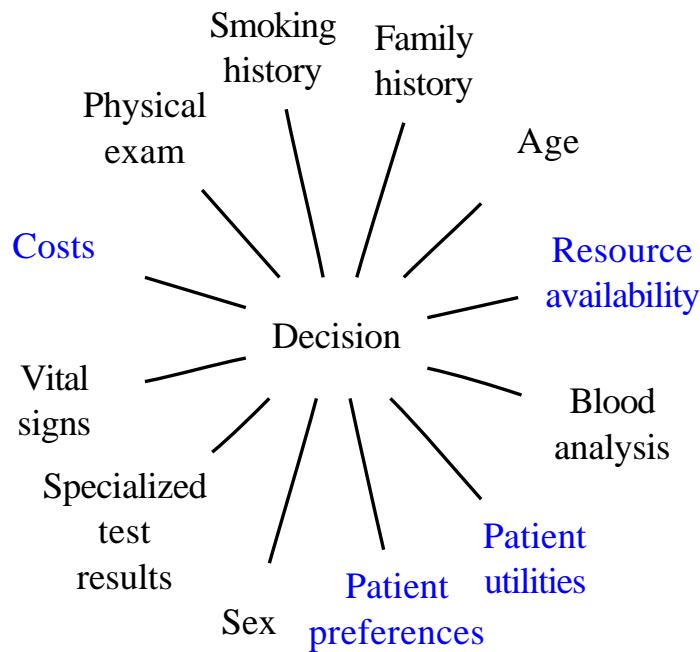
- Many studies attempt to classify patients as diseased/normal
- Given a reliable estimate of the probability of disease and the consequences of +/- one can make an optimal decision
- Consequences are known at the point of care, not by the authors; categorization **only** at point of care
- Continuous probabilities are self-contained, with their own “error rates”
- Middle probs. allow for “gray zone”, deferred decision

Patient	Prob[disease]	Decision	Prob[error]
1	0.03	normal	0.03
2	0.40	normal	0.40
3	0.75	disease	0.25

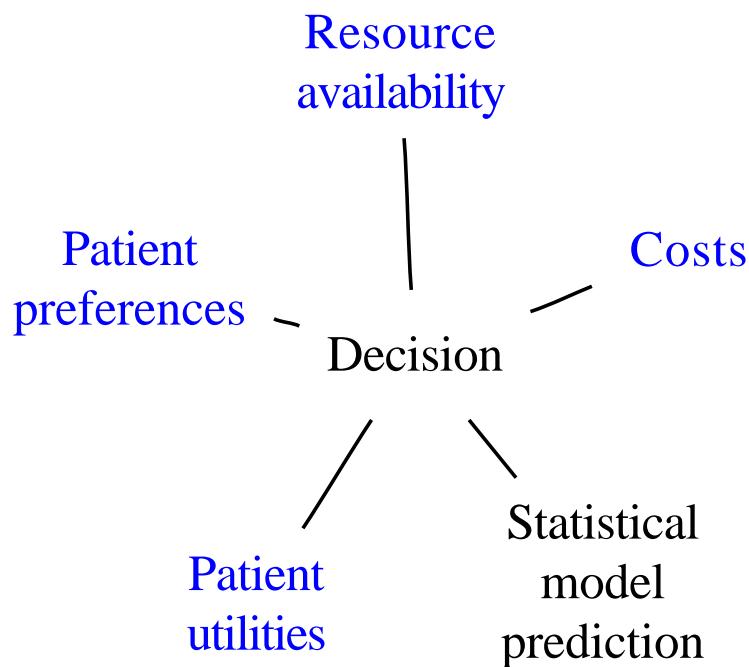
Note that much of diagnostic research seems to be aimed at making optimum decisions for groups of patients. The optimum decision for a group (if such a concept even has meaning) is not optimum for individuals in the group.

18.3.6

Components of Optimal Decisions



Statistical models reduce the dimensionality of the problem *but not to unity*



18.4

Problems with Classification of Predictions



- Feature selection / predictive model building requires choice of a scoring rule, e.g. correlation coefficient or proportion of correct classifications
- Prop. classified correctly is a discontinuous **improper scoring rule**
 - Maximized by bogus model (example below)
- Minimum information
 - low statistical power
 - high standard errors of regression coefficients
 - arbitrary to choice of cutoff on predicted risk
 - forces binary decision, does not yield a “gray zone” → more data needed
- Takes analyst to be provider of utility function and not the treating physician
- Sensitivity and specificity are also improper scoring rules

See bit.ly/risk-thresholds: Three Myths About Risk Thresholds for Prediction Models by Wynants *et al.*

18.4.1

Example: Damage Caused by Improper Scoring Rule

- Predicting probability of an event, e.g., $\text{Prob}[\text{disease}]$
- $N = 400$, 0.57 of subjects have disease
- Classify as diseased if $\text{prob.} > 0.5$

Model	<i>c</i> Index	χ^2	Proportion Correct
age	.592	10.5	.622
sex	.589	12.4	.588
age+sex	.639	22.8	.600
constant	.500	0.0	.573

Adjusted Odds Ratios:

age (IQR 58y:42y) 1.6 (0.95CL 1.2-2.0)

sex (f:m) 0.5 (0.95CL 0.3-0.7)

Test of sex effect adjusted for age ($22.8 - 10.5$):

$P = 0.0005$

Example where an improper accuracy score resulted in incorrect original analyses and incorrect re-analysis

Michiels, Koscielny, and Hill [65] used an improper accuracy score (proportion classified “correctly”) and claimed there was really no signal in all the published gene microarray studies they could analyze. This is true from the standpoint of repeating the original analyses (which also used improper accuracy scores) using multiple splits of the data, exposing the fallacy of using single data-splitting for validation. Aliferis et al. [3] used a semi-proper accuracy score (*c*-index) and they repeated 10-fold cross-validation 100 times instead of using highly volatile data splitting. They showed that the gene microarrays did indeed have predictive signals.^b

Michiels, Koscielny, and Hill [65]	Aliferis et al. [3]
% classified correctly	<i>c</i> -index
Single split-sample validation	Multiple repeats of 10-fold CV
Wrong tests (censoring, failure times)	Correct tests
5 of 7 published microarray studies had no signal	6 of 7 have signals

^bAliferis et al. [3] also used correct statistical models for time-to-event data that properly accounted for variable follow-up/censoring.

18.5

Value of Continuous Markers

- Avoid arbitrary cutpoints
- Better risk spectrum
- Provides gray zone
- Increases power/precision

18.5.1

Prognosis in Prostate Cancer

```

load('~/doc/Talks/infoAllergy/kattan.rda')
attach(kattan)
t    ← t.stg
gs   ← bx.glsn
psa  ← preop.psa
t12 ← t.stg %in% Cs(T1C,T2A,T2B,T2C)

s ← score.binary(t12 & gs ≤ 6 & psa < 10,
                  t12 & gs ≤ 6 & psa ≥ 10 & psa < 20,
                  t12 & gs == 7 & psa < 20,
                  (t12 & gs ≤ 6 & psa ≥ 20) | 
                  (t12 & gs ≥ 8 & psa < 20),
                  t12 & gs ≥ 7 & psa ≥ 20,
                  t.stg == 'T3')

levels(s) ← c('none', 'I', 'IIA', 'IIB', 'IIIA', 'IIIB', 'IIIC')
u ← is.na(psa + gs) | is.na(t.stg)
s[s == 'none'] ← NA
s ← s[drop=TRUE]
s3 ← s
levels(s3) ← c('I', 'II', 'III', 'III', 'III', 'III')
table(s3)

```

	I	II	III
1108	607	271	

```

units(time.event) ← 'month'
dd ← datadist(data.frame(psa, gs)); options(datadist='dd')
S ← Surv(time.event, event=='YES')
label(psa) ← 'PSA'; label(gs) ← 'Gleason Score'
f ← cph(S ~ rcs(sqrt(psa), 4), surv=TRUE, x=TRUE, y=TRUE)
p ← Predict(f, psa, time=24, fun=function(x) 1 - x)
h ← cph(S ~ s3, surv=TRUE)

```

```
z <- 1 - survest(h, times=24)$surv
```

```
ggplot(p, rdata=data.frame(psa), ylab='2-year Disease Recurrence Risk') +
  geom_hline(yintercept=unique(z), col='red', size=0.2)    # Fig. 18.3
```

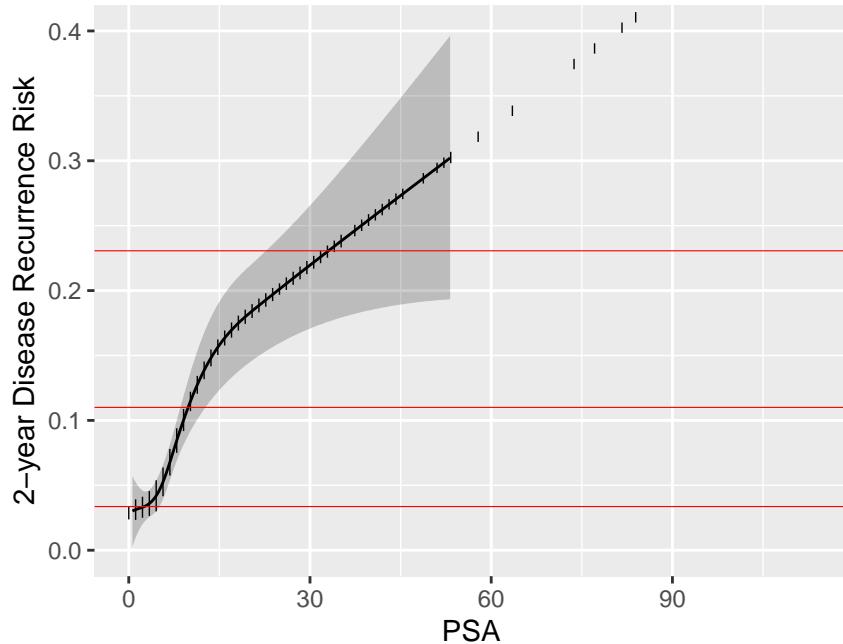


Figure 18.3: Relationship between post-op PSA level and 2-year recurrence risk. Horizontal lines represent the only prognoses provided by the new staging system. Data are courtesy of M Kattan from JNCI 98:715; 2006. Modification of AJCC staging by Roach *et al.* 2006.

Now examine the entire spectrum of estimated prognoses from variables models and from discontinuous staging systems.

```
f <- cph(S ~ rcs(sqrt(psa),4) + pol(gs,2), surv=TRUE)
g <- function(form, lab) {
  f <- cph(form, surv=TRUE, subset=!u)
  cat(lab, '\n'); print(coef(f))
  s <- f$stats
  cat('N:', s['Obs'], '\tL.R.:', round(s['Model L.R.'],1),
      '\td.f.:', s['d.f.'], '\n\n')
  prob24 <- 1 - survest(f, times=24)$surv
  prn(sum(!is.na(prob24)))
  p2 <- c(p2, prob24[2]) # save est. prognosis for one subject
  p1936 <- c(p1936, prob24[1936])
  C <- rcorr.cens(1-prob24, S[!u,])['C Index']
  data.frame(model=lab, chisq=s['Model L.R.'], d.f.=s['d.f.'],
             C=C, prognosis=prob24)
}
p2 <- p1936 <- NULL
w <- g(S ~ t.stg, 'Old Stage')
```

Old Stage
t.stg=T2A t.stg=T2B t.stg=T2C t.stg=T3
0.2791987 1.2377218 1.0626197 1.7681393
N: 1978 L.R.: 70.5 d.f.: 4

```
sum(!is.na(prob24))
```

```
[1] 1978
```

```
w ← rbind(w, g(S ~ s3, 'New Stage'))
```

```
New Stage
  s3=II    s3=III
1.225296 1.990355
N: 1978      L.R.: 135.8      d.f.: 2
```

```
sum(!is.na(prob24))
```

```
[1] 1978
```

```
w ← rbind(w, g(S ~ s, 'New Stage, 6 Levels'))
```

```
New Stage, 6 Levels
  s=IIA    s=IIB    s=IIIA   s=IIIB   s=IIIC
1.181824 1.248864 1.829265 2.410810 1.954420
N: 1978      L.R.: 140.3      d.f.: 5
```

```
sum(!is.na(prob24))
```

```
[1] 1978
```

```
w ← rbind(w, g(S ~ pol(gs, 2), 'Gleason'))
```

```
Gleason
  gs        gs^2
-0.42563792 0.07857747
N: 1978      L.R.: 90.3      d.f.: 2
```

```
sum(!is.na(prob24))
```

```
[1] 1978
```

```
w ← rbind(w, g(S ~ rcs(sqrt(psa), 4), 'PSA'))
```

```
PSA
  psa        psa'       psa''
-0.09621478 4.07465107 -14.86458188
N: 1978      L.R.: 95.3      d.f.: 3
```

```
sum(!is.na(prob24))
```

```
[1] 1978
```

```
w ← rbind(w, g(S ~ rcs(sqrt(psa), 4) + pol(gs, 2), 'PSA+Gleason'))
```

```
PSA+Gleason
      psa          psa'         psa''          gs          gs^2
-0.11703664  3.37768454 -12.04890937 -0.20429572  0.05458832
N: 1978          L.R.: 160.2       d.f.: 5
```

```
sum(!is.na(prob24))

[1] 1978
```

```
w ← rbind(w, g(S ~ rcs(sqrt(psa), 4) + pol(gs, 2) + t.stg,
  'PSA+Gleason+Old Stage'))
```

```
PSA+Gleason+Old Stage
      psa          psa'         psa''          gs          gs^2    t.stg=T2A
0.12361025  2.26959366 -8.62949512 -0.01467426  0.03511191  0.27334309
  t.stg=T2B    t.stg=T2C    t.stg=T3
0.93943683  0.69083735  1.07508642
N: 1978          L.R.: 187       d.f.: 9
```

```
sum(!is.na(prob24))

[1] 1978
```

```
w$z ← paste(w$model, '\n',
  'X2-d.f.=', round(w$chisq-w$d.f.),
  ' C=', sprintf("%.2f", w$C), sep=' ')
w$z ← with(w, factor(z, unique(z)))
require(lattice)
stripplot(z ~ prognosis, data=w, lwd=1.5,      # Fig. 18.4
  panel=function(x, y, ...) {
    llines(p2, 1:7, col=gray(.6))
    ## llines(p1936, 1:7, col=gray(.8), lwd=2)
    ## panel.stripplot(x, y, ..., jitter.data=TRUE, cex=.5)
    for(iy in unique(unclass(y))) {
      s ← unclass(y)==iy
      histSpike(x[s], y=rep(iy, sum(s)), add=TRUE, grid=TRUE)
    }
    panel.abline(v=0, col=gray(.7))
  },
  xlab='Predicted 2-year\nDisease Recurrence Probability')
```

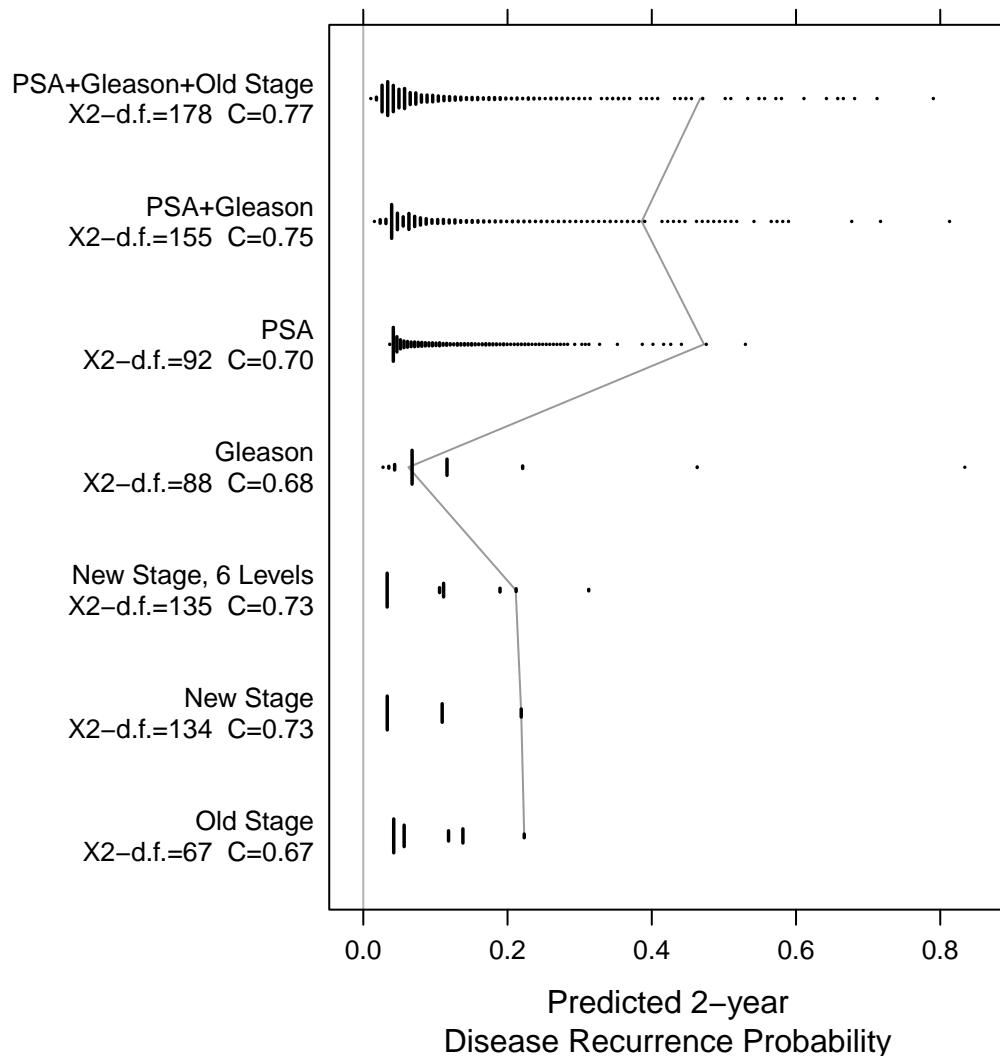


Figure 18.4: Prognostic spectrum from various models with model χ^2 - d.f., and generalized c -index. The mostly vertical segmented line connects different prognostic estimates for the same man.

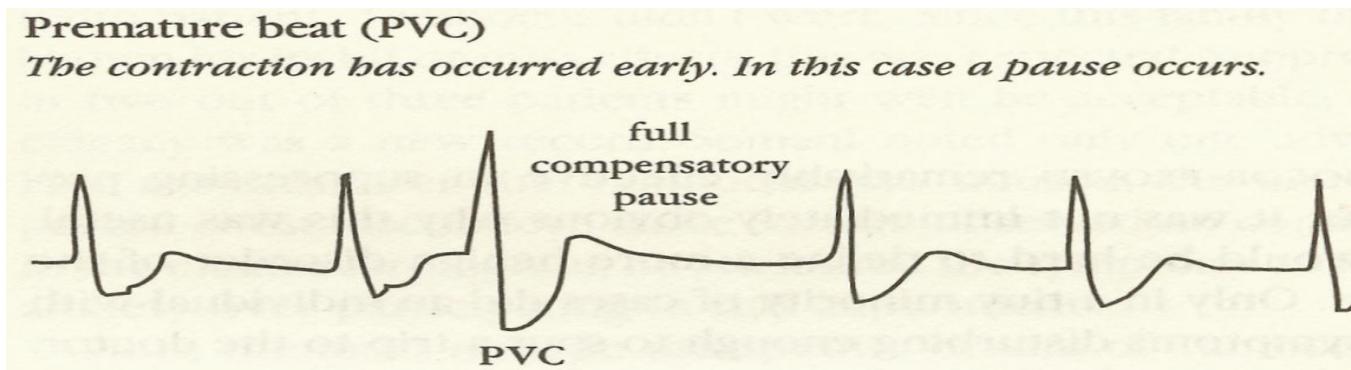
18.6

Harm from Ignoring Information

18.6.1

Case Study: Cardiac Anti-arrhythmic Drugs

- Premature ventricular contractions were observed in patients surviving acute myocardial infarction
- Frequent PVCs ↑ incidence of sudden death



Moore [68], p. 46

Arrhythmia Suppression Hypothesis

Any prophylactic program against sudden death must involve the use of anti-arrhythmic drugs to subdue ventricular premature complexes.

Bernard Lown
Widely accepted by 1978

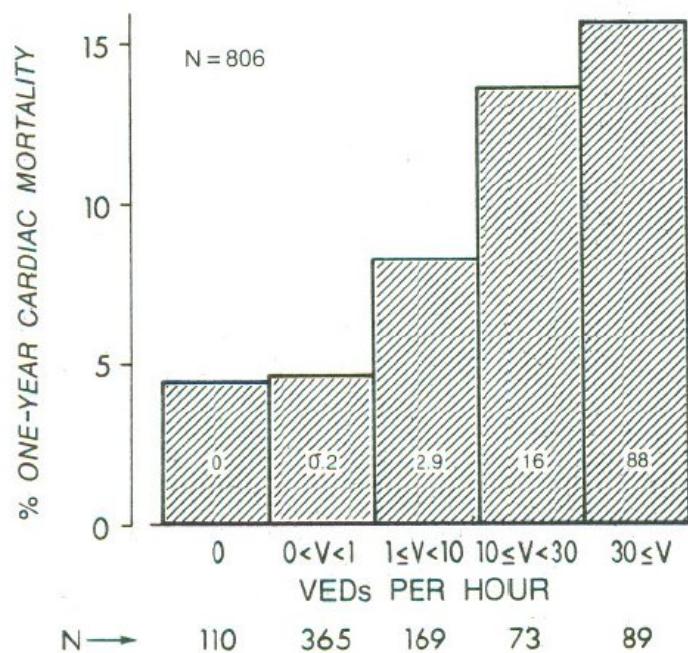


Figure 2. Cardiac Mortality Rate in Five Categories for Frequency of Ventricular Ectopic Depolarizations (VEDs) Determined by 24-Hour Holter Recording before Discharge.

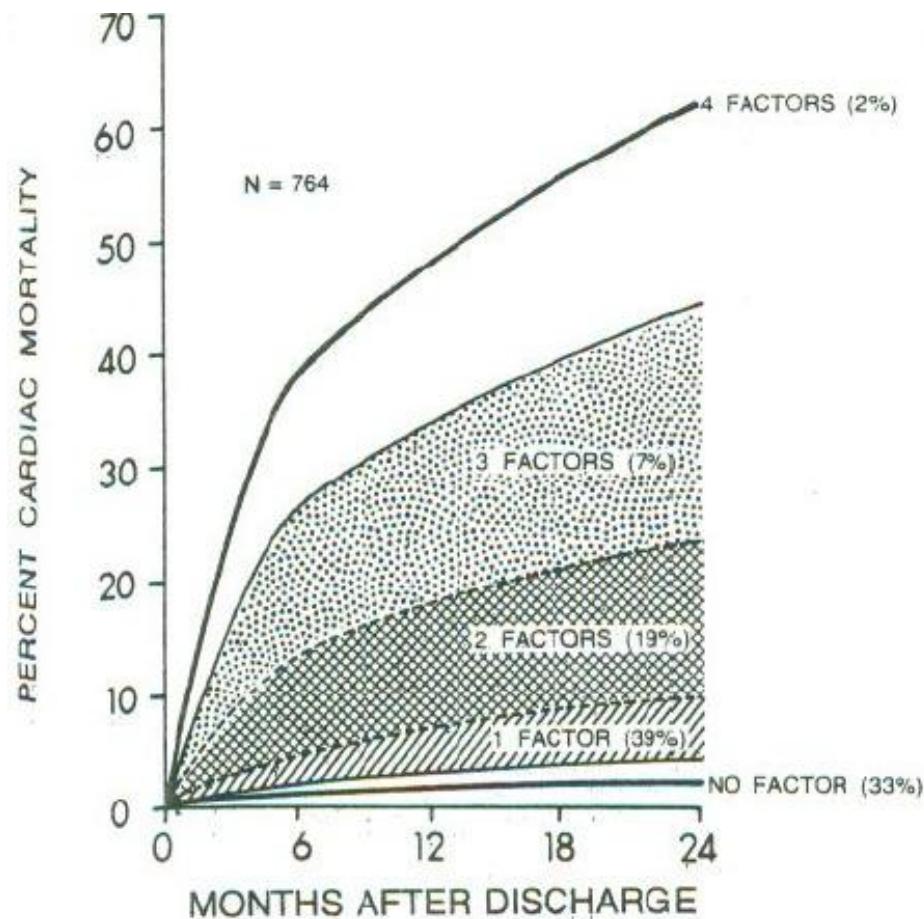
N denotes the number of patients in the total population and in each category. Of 819 patients with Holter recordings, 13 were lost to follow-up during the first year after hospitalization. The numbers within each of the boxes denote the median frequency of ventricular ectopy.

Moore [68], p. 49;⁶⁹

Are PVCs independent risk factors for sudden cardiac death?

Researchers developed a 4-variable model for prognosis after acute MI

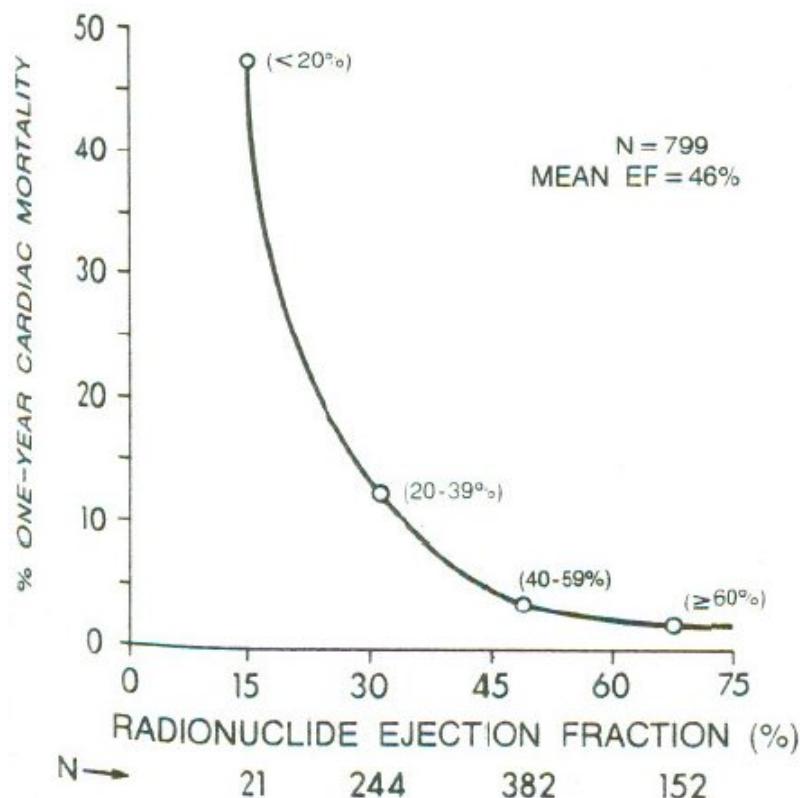
- left ventricular ejection fraction (EF) < 0.4
- PVCs > 10/hr
- Lung rales
- Heart failure class II, III, IV



Multicenter Postinfarction Research Group [69]

Dichotomania Caused Severe Problems

- EF alone provides same prognostic spectrum as the researchers' model
- Did not adjust for EF!; PVCs ↑ when $EF < 0.2$
- Arrhythmias prognostic in isolation, not after adjustment for continuous EF and anatomic variables
- Arrhythmias predicted by local contraction abnorm., then global function (EF)



Multicenter Postinfarction Research Group [69]; Califf et al. [14]

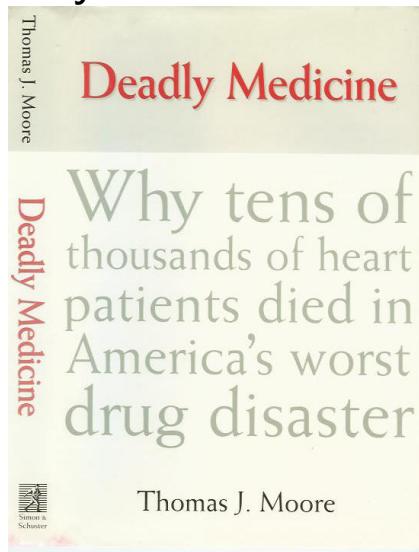
18.6.2

CAST: Cardiac Arrhythmia Suppression Trial

- Randomized placebo, moricizine, and Class IC anti-arrhythmic drugs flecainide and encainide
- Cardiologists: unethical to randomize to placebo
- Placebo group included after vigorous argument
- Tests design as one-tailed; did not entertain possibility of harm
- Data and Safety Monitoring Board recommended early termination of flecainide and encainide arms
- Deaths $\frac{56}{730}$ drug, $\frac{22}{725}$ placebo, RR 2.5

Investigators [49]

Conclusions: Class I Anti-Arrhythmics



Estimate of excess deaths from Class I anti-arrhythmic drugs: 24,000–69,000
Estimate of excess deaths from Vioxx: 27,000–55,000

Arrhythmia suppression hypothesis refuted; PVCs merely indicators of underlying, permanent damage

Moore [68], pp. 289,49; D Graham, FDA

18.7

Case Study in Faulty Dichotomization of a Clinical Outcome: Statistical and Ethical Concerns in Clinical Trials for Crohn's Disease

18.7.1

Background

Many clinical trials are underway for studying treatments for Crohn's disease. The primary endpoint for these studies is a discontinuous, information-losing transformation of the Crohn's Disease Activity Index (CDAI)⁸, which was developed in 1976 by using an exploratory stepwise regression method to predict four levels of clinicians' impressions of patients' current status^c. The first level ("very well") was assumed to indicate the patient was in remission. The model was overfitted and was not validated. The model's coefficients were scaled and rounded, resulting in the following scoring system (see <http://www.ibdjohn.com/cdai>).

^cOrdinary least squares regression was used for the ordinal response variable. The levels of the response were assumed to be equally spaced in severity on a numerical scale of 1, 3, 5, 7 with no justification.

Crohn's Disease Activity Index
 $\text{CDAI} = 2 \times 1 + 5 \times 2 + 7 \times 3 + 20 \times 4 + 30 \times 5 + 10 \times 6 + 6 \times 7 + (\text{weight factor})_8$

The purpose of this crohn's disease activity index (CDAI) calculator is to gauge the progress or lack of progress for people with crohn's disease. The reference article says "generally speaking, CDAI scores below 150 indicate a better prognosis than higher scores." (See Reference at bottom).

However, since the original study, other researchers use a 'subjective value' of 200 to 250. Therefore, this just reinforces the fact that the purpose is to gauge Your Progress i.e. compare readings from one week to the next to determine if you are getting better or worst. Bottom line is that you need to use the CDAI on a regular basis and view it as a personal gauge. Watch for changes in your score (your gauge).

This 'indicator' does NOT predict the outcome of the disease. Crohn's disease conditions vary for each patient. This calculator is only a 'gauge' of progress i.e. not a prognosis tool!

1. Number of liquid or very soft stools in one week

Input:
 Total For One Week

2. Sum of seven daily abdominal pain ratings:
 (0=none, 1=mild, 2=moderate, 3=severe)

Overall Rating: 0 1 2 3

3. Sum of seven daily ratings of general well-being:
 (0=well, 1=slightly below par, 2=poor, 3=very poor, 4=terrible)

Rating: 0 1 2 3 4

4. Symptoms or findings presumed related to Crohn's disease
 Select each set corresponding to patient's symptoms:

arthritis or arthralgia
 iritis or uveitis
 erythema nodosum, pyoderma gangrenosum, aphthous stomatitis
 anal fissure, fistula or perirectal abscess
 other bowel-related fistula
 febrile (fever) episode over 100 degrees during past week

5. Taking Lomotil or opiates for diarrhea

No Yes

6. Abnormal mass
 0=none; 0.4=questionable; 1=present

None Questionable Present

7. Hematocrit [(Typical - Current) x 6]
 Normal average: For Male = 47 For Female = 42
VIP: Skip this section if typical and current are unknown.

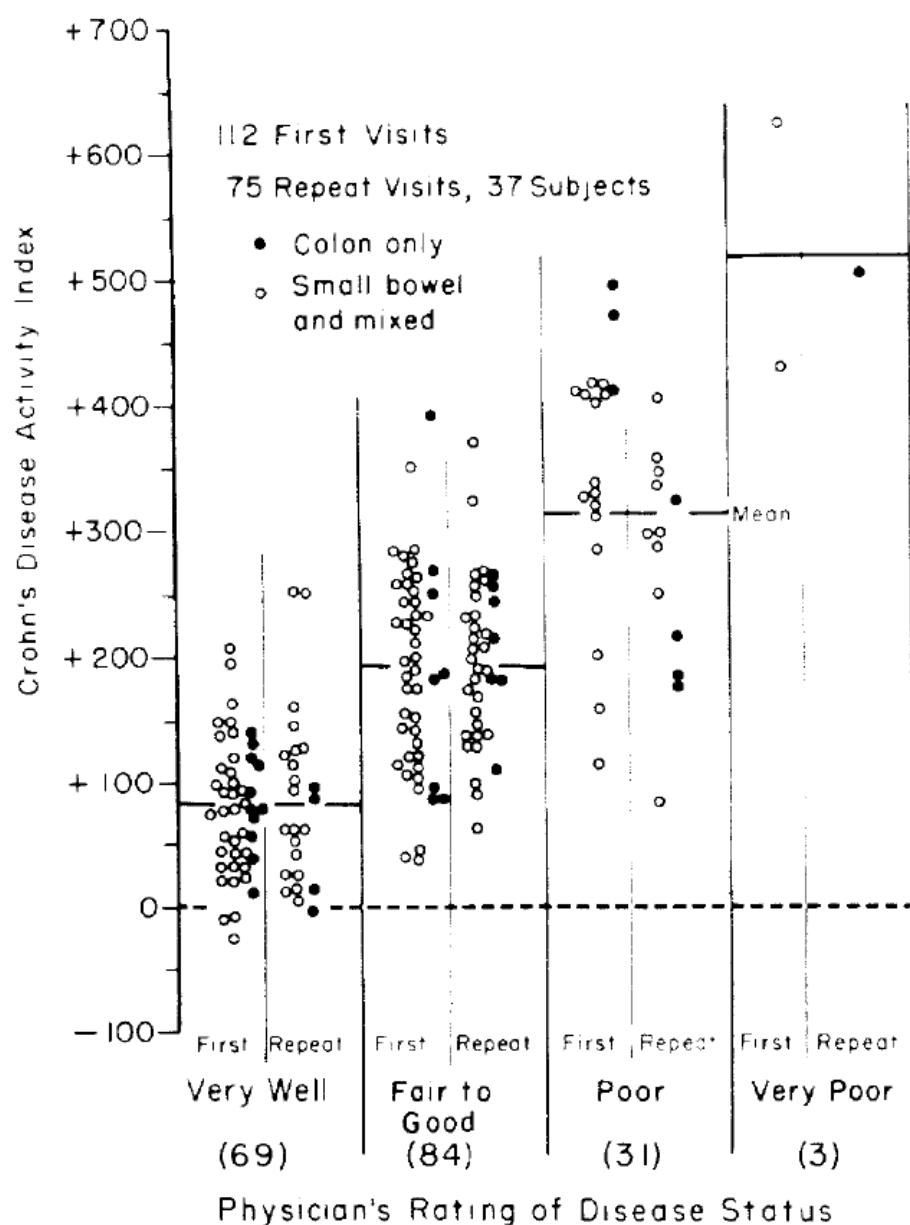
Enter 'YOUR' typical value and the current value
 If you want to include this calculation
 Male Female
 Enter Typical = Current =

8. $100 \times [(\text{standard weight-actual body weight}) / \text{standard weight}]$

Recommendation:
 Skip this section unless weight 'changes related' to crohn's are known !
 The purpose is to check for weight change = change in conditions.
 For example: Weight loss such as that caused by dehydration.

Standard Weight
 Actual Weight
 Submit/Calculate

The original authors plotted the predicted scores against the four clinical categories as shown below.



The authors arbitrarily assigned a cutoff of 150, below which indicates "remission."^d It can be seen that "remission" includes a good number of patients actually classified as "fair to good" or "poor." A cutoff only exists when there is a break in the distribution of scores. As an example, data were simulated from a population in which every patient having a score below 100 had a probability of response of 0.2 and every patient having a score above 100 had a probability of response of 0.8. Histograms showing the distributions of non-responders (just above the x -axis) and responders (at the top of the graph) appear in the figure below. A flexibly fitted logistic regression model relating

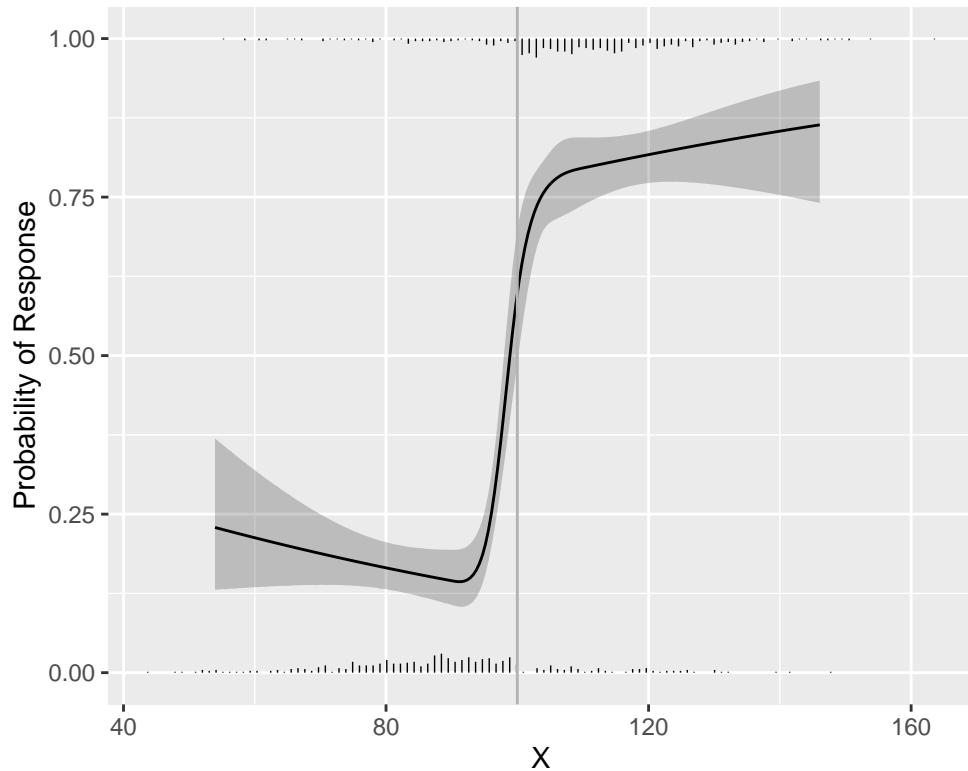
^dHowever, the authors intended for CDAI to be used on a continuum: "...a numerical index was needed, the numerical value of which would be proportional to degree of illness ... it could be used as the principal measure of response to the therapy under trial ... the CDAI appears to meet those needs. ... The data presented ... is an accurate numerical expression of the physician's over-all assessment of degree of illness in a large group of patients ... we believe that it should be useful to all physicians who treat Crohn's disease as a method of assessing patient progress."

observed scores to actual response status is shown, along with 0.95 confidence intervals for the fit.

```
require(rms)
set.seed(4)
n <- 900
X <- rnorm(n, 100, 20)
dd <- datadist(X); options(datadist='dd')

p <- ifelse(X < 100, .2, .8)
y <- ifelse(runif(n) ≤ p, 1, 0)

f <- lrm(y ~ rcs(X, c(90,95,100,105,110)))
hs <- function(yval, side)
  histSpikeg(yhat ~ X, data=subset(data.frame(X, y), y == yval),
             side = side, ylim = c(0, 1),
             frac = function(f) .03 * f / max(f))
ggplot(Predict(f, fun=plogis), ylab='Probability of Response') +
  hs(0, 1) + hs(1, 3) + geom_vline(xintercept=100, col=gray(.7))
```



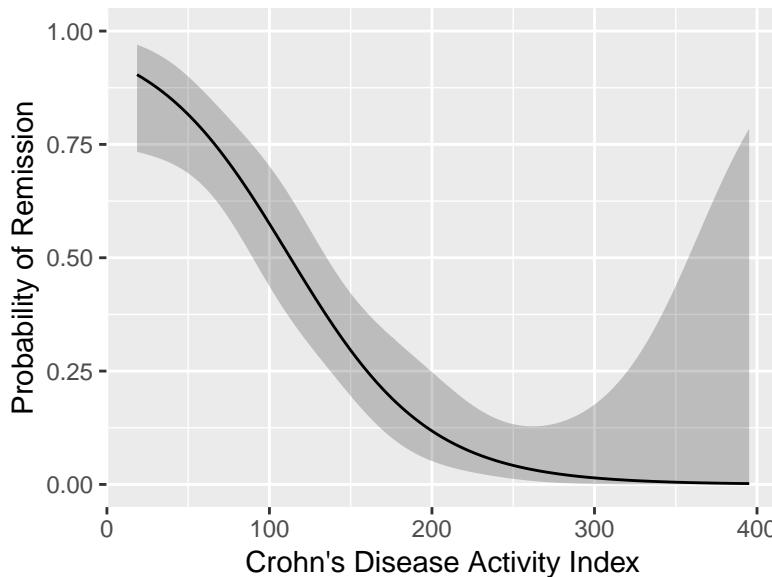
One can see that the fitted curves justify the use of a cut-point of 100. However, the original scores from the development of CDAI do not justify the existence of a cutoff. The fitted logistic model used to relate “very well” to the other three categories is shown below.

```
# Points from published graph were defined in code not printed
g <- trunc(d$x)
g <- factor(g, 0:3, c('very well', 'fair to good', 'poor', 'very poor'))
remiss <- 1 * (g == 'very well')
CDAI <- d$y
label(CDAI) <- "Crohn's Disease Activity Index"
```

```

label(remiss) ← 'Remission'
dd ← datadist(CDAI,remiss); options(datadist='dd')
f ← lrm(remiss ~ rcs(CDAI,4))
ggplot(Predict(f, fun=plogis), ylab='Probability of Remission')

```



It is readily seen that no cutoff exists, and one would have to be below CDAI of 100 for the probability of remission to fall below even 0.5. The probability does not exceed 0.9 until the score falls below 25. Thus there is no clinical justification for the 150 cut-point.

18.7.2

Loss of Information from Using Cut-points

The statistical analysis plan in the Crohn's disease protocols specify that efficacy will be judged by comparing two proportions after classifying patients' CDAs as above or below the cutoff of 150. Even if one could justify a certain cutoff from the data, the use of the cutoff is usually not warranted. This is because of the huge loss of statistical efficiency, precision, and power from dichotomizing continuous variables as discussed in more detail in Section 18.3.4. If one were forced to dichotomize a continuous response Y , the cut-point that loses the least efficiency is the population median of Y combining treatment groups. That implies a statistical efficiency of $\frac{2}{\pi}$ or 0.637 when compared to the efficient two-sample t -test if the data are normally distributed^e. In other words, the optimum cut-point would require studying 158 patients after dichotomizing the

^eNote that the efficiency of the Wilcoxon test compared to the t -test is $\frac{3}{\pi}$ and the efficiency of the sign test compared to the t -test is $\frac{2}{\pi}$. Had analysis of covariance been used instead of a simple two-group comparison, the baseline level of CDAI could have been adjusted for as a covariate. This would have increased the power of the continuous scale approach to even higher levels.

response variable to get the same power as analyzing the continuous response variable in 100 patients.

18.7.3

Ethical Concerns and Summary

The CDAI was based on a sloppily-fit regression model predicting a subjective clinical impression. Then a cutoff of 150 was used to classify patients as in remission or not. The choice of this cutoff is in opposition to the data used to support it. The data show that one must have CDAI below 100 to have a chance of remission of only 0.5. Hence the use of $\text{CDAI} < 150$ as a clinical endpoint was based on a faulty premise that apparently has never been investigated in the Crohn's disease research community. CDAI can easily be analyzed as a continuous variable, preserving all of the power of the statistical test for efficacy (e.g., two-sample t -test). The results of the t -test can readily be translated to estimate any clinical "success probability" of interest, using efficient maximum likelihood estimators^f

There are substantial ethical questions that ought to be addressed when statistical power is wasted:

1. Patients are not consenting to be put at risk for a trial that doesn't yield valid results.
2. A rigorous scientific approach is necessary in order to allow enrollment of individuals as subjects in research.
3. Investigators are obligated to reduce the number of subjects exposed to harm and the amount of harm to which each subject is exposed.

It is not known whether investigators are receiving per-patient payments for studies in which sample size is inflated by dichotomizing CDAI.

^fGiven \bar{x} and s as estimates of μ and σ , the estimate of the probability that $\text{CDAI} < 150$ is simply $\Phi(\frac{150 - \bar{x}}{s})$, where Φ is the cumulative distribution function of the standard normal distribution. For example, if the observed mean were 150, we would estimate the probability of remission to be 0.5.

18.8

Information May Sometimes Be Costly

When the Missionaries arrived, the Africans had the Land and the Missionaries had the Bible. They taught how to pray with our eyes closed. When we opened them, they had the land and we had the Bible.

Information itself has a liberal bias.

Jomo Kenyatta, founding father of Kenya; also attributed to Desmond Tutu

The Colbert Report
2006-11-28

Information Allergy

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University
Office of Biostatistics
FDA CDER

Information allergy is defined as (1) refusing to obtain key information needed to make a sound decision, or (2) ignoring important available information. The latter problem is epidemic in biomedical and epidemiologic research and in clinical practice. Examples include

- ignoring some of the information in confounding variables that would explain away the effect of characteristics such as dietary habits
- ignoring probabilities and “gray zones” in genomics and proteomics research, making arbitrary classifications of patients in such a way that leads to poor validation of gene and protein patterns
- failure to grasp probabilistic diagnosis and patient-specific costs of incorrect decisions, thus making arbitrary diagnoses and placing the analyst in the role of the bedside decision maker
- classifying patient risk factors and biomarkers into arbitrary “high/low” groups, ignoring the full spectrum of values
- touting the prognostic value of a new biomarker, ignoring basic clinical information that may be even more predictive
- using weak and somewhat arbitrary clinical staging systems resulting from a fear of continuous measurements
- ignoring patient spectrum in estimating the benefit of a treatment

Examples of such problems will be discussed, concluding with an examination of how information–losing cardiac arrhythmia research may have contributed to the deaths of thousands of patients.

Chapter 19

Diagnosis



Medical diagnostic research, as usually practiced, is prone to bias and even more importantly to yielding information that is not useful to patients or physicians and sometimes overstates the value of diagnostic tests. Important sources of these problems are conditioning on the wrong statistical information, reversing the flow of time, and categorization of inherently continuous test outputs and disease severity. It will be shown that sensitivity and specificity are not properties of tests in the usual sense of the word, and that they were never natural choices for describing test performance. This implies that ROC curves are unhelpful (although areas under them are sometimes useful). So is categorical thinking.

This chapter outlines the many advantages of diagnostic risk modeling, showing how pre- and post-test diagnostic models give rise to clinically useful displays of pre-test vs. post-test probabilities that themselves quantify diagnostic utility in a way that is useful to patients, physicians, and diagnostic device makers. And unlike sensitivity and specificity, post-test probabilities are immune to certain biases, including workup bias.

Case-control studies use a design where sampling is done on final disease status and patient exposures are “along for the ride.” In other words, one conditions on the outcome and considers the distribution of exposures using outcome-dependent sampling. Sensitivity and specificity are useful for proof-of-concept case-control studies because sensitivity and specificity also condition on the final diagnosis. The use of sensitivity and specificity in prospective cohort studies is the mathematical equivalent of making three left turns in order to turn right. Like the complex adjustments needed for *P*-values when doing sequential trials, sensitivity and specificity require complex adjustments for workup bias just because of their backward consideration of time and information. In

Most of this material is from “Direct Measures of Diagnostic Utility Based on Diagnostic Risk Models” by FE Harrell presented at the FDA Public Workshop on Study Methodology for Diagnostics in the Postmarket Setting, 2011-05-12.

a cohort study one can use a vanilla regression model to estimate the probability of a final diagnostic outcome given patient characteristics and diagnostic test results.

19.1

Problems with Traditional Indexes of Diagnostic Utility

$$\text{sensitivity} = \text{Prob}[T^+|D^+]$$

$$\text{specificity} = \text{Prob}[T^-|D^-]$$

$$\text{Prob}[D^+|T^+] = \frac{\text{sens} \times \text{prev}}{\text{sens} \times \text{prev} + (1-\text{spec}) \times (1-\text{prev})}$$

Problems:

- Diagnosis forced to be binary
- Test forced to be binary
- Sensitivity and specificity are in backwards time order
- Confuse decision making for groups vs. individuals
- Inadequate utilization of pre-test information
- Dichotomization of continuous variables in general

Example: BI-RADS Score in Mammography

Does Category 4 Make **Any** Sense?

	Diagnosis	Number of Criteria^a
0	Incomplete	Your mammogram or ultrasound didn't give the radiologist enough information to make a clear diagnosis; follow-up imaging is necessary
1	Negative	There is nothing to comment on; routine screening recommended
2	Benign	A definite benign finding; routine screening recommended
3	Probably Benign	Findings that have a high probability of being benign (> 98%); six-month short interval follow-up
4	Suspicious Abnormality	Not characteristic of breast cancer, but reasonable probability of being malignant (3 to 94%); biopsy should be considered
5	Highly Suspicious of Malignancy	Lesion that has a high probability of being malignant ($\geq 95\%$); take appropriate action
6	Known Biopsy Proven Malignancy	Lesions known to be malignant that are being imaged prior to definitive treatment; assure that treatment is completed

How to Reduce False Positives and Negatives?

- Do away with “positive” and “negative”
- Provide risk estimates
- Defer decision to decision maker
- Risks have self-contained error rates
- Risk of 0.2 \rightarrow Prob[error]=.2 if don’t treat
- Risk of 0.8 \rightarrow Prob[error]=.2 if treat

See <http://thehealthcareblog.com/blog/2015/12/01/rethinking-about-diagnostic-tests-there-is-nothing-positive-or-negative-about-a-test-result> for a nice article on the subject.

Binary Diagnosis is Problematic Anyway

The act of diagnosis requires that patients be placed in a binary category of either having or not having a certain disease. Accordingly, the diseases of particular concern for industrialized countries—such as type 2 diabetes, obesity, or depression—require that a somewhat arbitrary cut-point be chosen on a continuous scale of measurement (for example, a fasting glucose level > 6.9 mmol/L [> 125 mg/dL] for type 2 diabetes). These cut-points do not adequately reflect disease biology, may inappropriately treat patients on either side of the cut-point as 2 homogeneous risk groups, fail to incorporate other risk factors, and are invariable to patient preference.

Vickers, Basch, and Kattan [109]

Newman and Kohn⁷⁴ have a strong section about the problems with considering diagnosis to be binary.

Back to Sensitivity and Specificity

- Backwards time order
- Irrelevant to both physician and patient
- Improper discontinuous scoring rules^b
- Are not test characteristics
 - Are characteristics of the test **and** patients
- **Not constant**; vary with patient characteristics
 - Sensitivity ↑ with any covariate related to disease severity if diagnosis is dichotomized
- Require adjustment for workup bias
 - Diagnostic risk models do not; only suffer from under-representation
- Good for proof of concept of a diagnostic method in a case-control study; not useful for utility

^bThey are optimized by not correctly estimating risk of disease.

Hlatky et al. [45];⁶⁷ Moons and Harrell [66]; Gneiting and Raftery [35]

Sensitivity of Exercise ECG for Diagnosing CAD

Age (years)	Sensitivity
< 40	0.56
40–49	0.65
50–59	0.74
≥ 60	0.84

Sex	
male	0.72
female	0.57

# Diseased CAs	
1	0.48
2	0.68
3	0.85

Hlatky et al. [45]. See also Janssens 2005

19.2

Problems with ROC Curves and Cutoffs

...statistics such as the AUC are not especially relevant to someone who must make a decision about a particular x_cROC curves lack or obscure several quantities that are necessary for evaluating the operational effectiveness of diagnostic tests. ...ROC curves were first used to check how radio *receivers* (like radar receivers) operated over a range of frequencies. ...This is not how most ROC curves are used now, particularly in medicine. The receiver of a diagnostic measurement ...wants to make a decision based on some x_c , and is not especially interested in how well he would have done had he used some different cutoff. Briggs and Zaretzki [12]

In the discussion to this paper, David Hand states “when integrating to yield the overall AUC measure, it is necessary to decide what weight to give each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data. This is nonsensical. The relative importance of misclassifying a case as a non-case, compared to the reverse, cannot come from the data itself. It must come externally, from considerations of the severity one attaches to the different kinds of misclassifications.”

19.3

Optimum Individual Decision Making and Forward Risk

- Minimize expected loss/cost/disutility
- Uses
 - utility function (e.g., inverse of cost of missing a diagnosis, cost of over-treatment if disease is absent)
 - probability of disease

d = decision, o = outcome

Utility for outcome o = $U(o)$

Expected utility of decision d = $U(d) = \int p(o|d)U(o)do$

d_{Opt} = d maximizing $U(d)$

The steps for determining the optimum decision are:

1. Predict the outcome o for every decision d
2. Assign a utility $U(o)$ to every outcome o
3. Find the decision d that maximizes the expected utility

See

- <https://bit.ly/datamethods-dm>
- http://en.wikipedia.org/wiki/Optimal_decision
- <http://www.statsathome.com/2017/10/12/bayesian-decision-theory-made-ridiculously-simple>
- https://twitter.com/jim_savage_/status/989524417649807360
- Govers et al.³⁶

- <https://stats.stackexchange.com/questions/368949>

See [this NY Times article](#) about decision theory.

19.4

Diagnostic Risk Modeling

Assuming (Atypical) Binary Disease Status

Y 1:diseased, 0:normal

X vector of subject characteristics (e.g., demographics, risk factors, symptoms)

T vector of test (biomarker, ...) outputs

α intercept

β vector of coefficients of X

γ vector of coefficients of T

$$\text{pre}(X) = \text{Prob}[Y = 1|X] = \frac{1}{1+\exp[-(\alpha^* + \beta^* X)]}$$

$$\text{post}(X, T) = \text{Prob}[Y = 1|X, T] = \frac{1}{1+\exp[-(\alpha + \beta X + \gamma T)]}$$

Note: Proportional odds model extends to ordinal disease severity Y .

19.4.1

Example Diagnostic Models

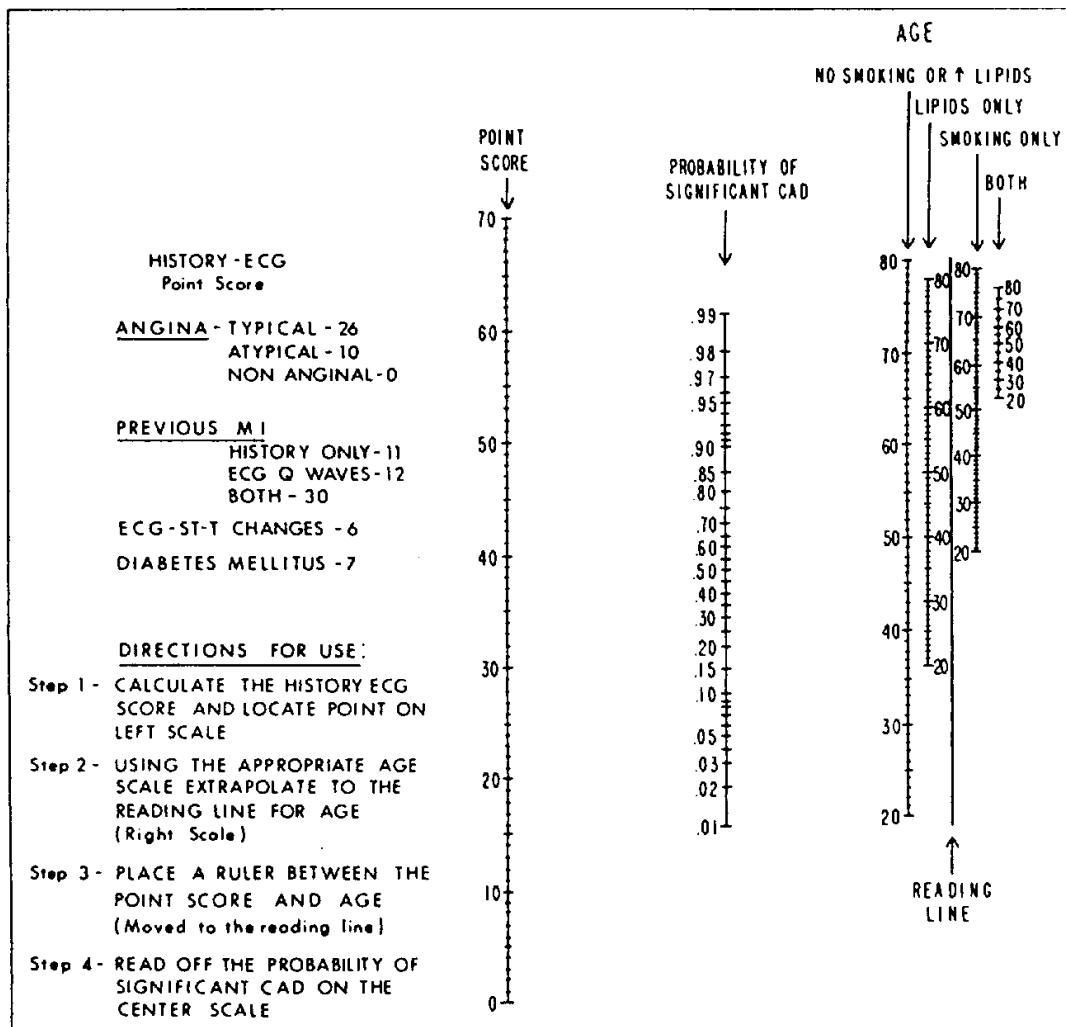


Figure 19.1: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. ECG = electrocardiographic; MI = myocardial infarction.⁸² Reprinted from American Journal of Medicine, Vol. 75, D.B. Pryor *et al.*, "Estimating the likelihood of significant coronary artery disease," p. 778, Copyright 1983, with permission from Excerpta Medica, Inc.

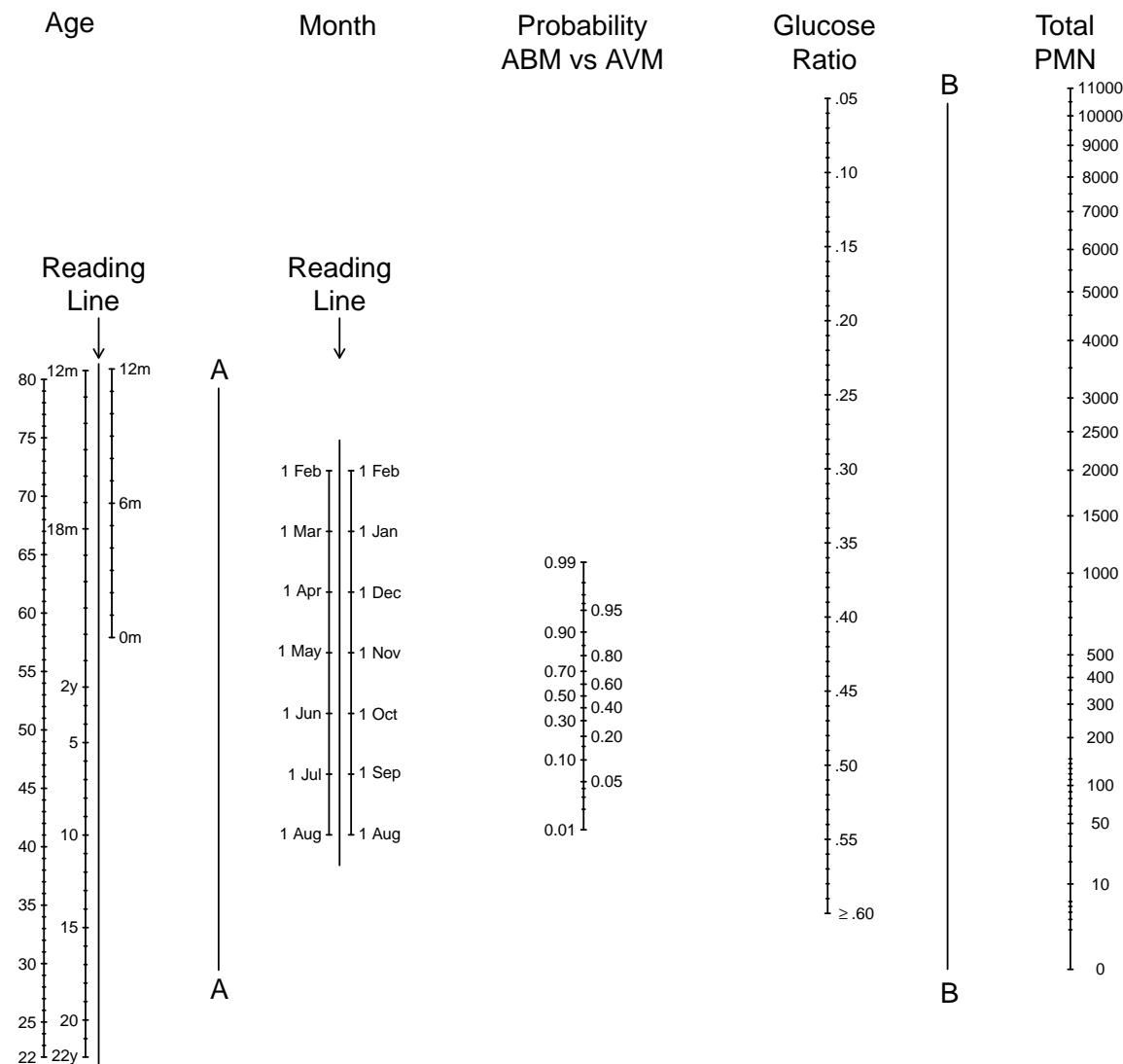


Figure 19.2: Nomogram for estimating probability of bacterial (ABM) versus viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerebrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM versus AVM. From Spanos, Harrell, and Durack [98]

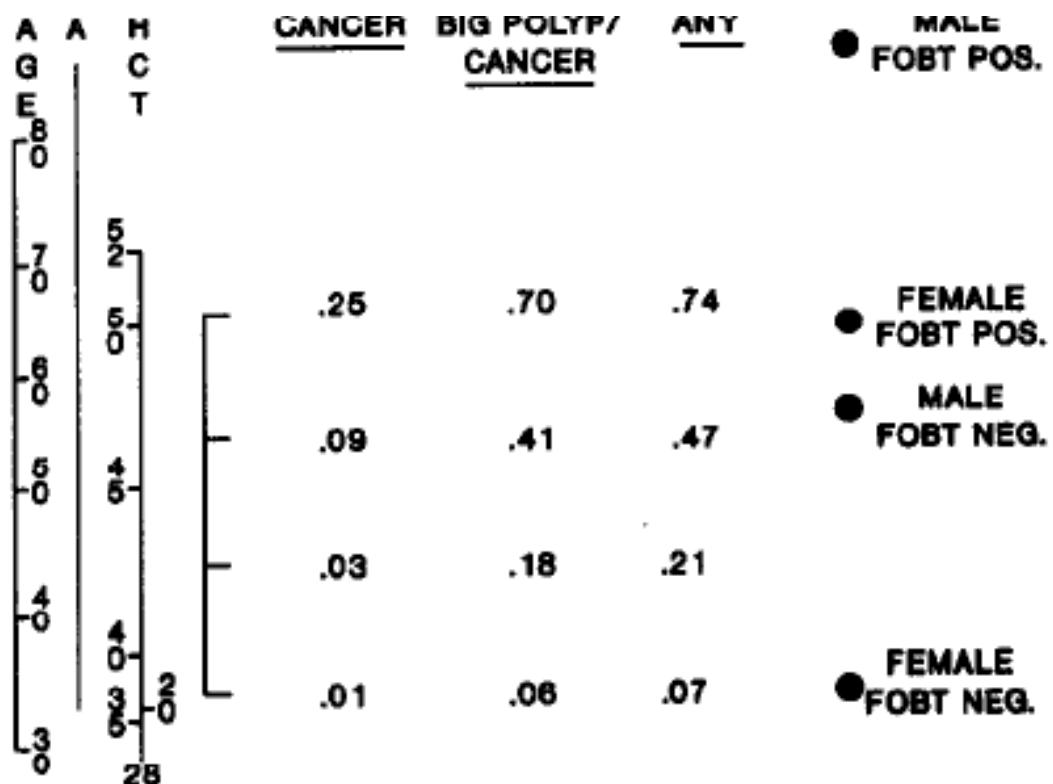


Fig. 3. A nomogram for estimating the likelihood of any colorectal neoplasia, an adenomatous polyp ≥ 5 mm or cancer. *Directions for use:* (1) If diarrhea or pain is the sole

Figure 19.3: Proportional odds ordinal logistic model for ordinal diagnostic classes from Brazer et al. [11]

19.5

Assessing Diagnostic Yield

19.5.1

Absolute Yield

Pencina et al. [80]: Absolute incremental information in a new set of markers

Consider change in predicted risk when add new variables

Average increase in risk of disease when disease present

+

Average decrease in risk of disease when disease absent

Formal Test of Added Absolute and Relative Information

Likelihood ratio χ^2 test of partial association of new markers, adjusted for old markers

19.5.2

Assessing Relative Diagnostic Yield

- Variation in relative log odds of disease = $T\hat{\gamma}$, holding X constant
- Summarize with Gini's mean difference or inter-quartile range, then anti-log
- E.g.: the typical modification of pre-test odds of disease is by a factor of 3.4

Gini's mean difference = mean absolute difference between any pair of values

See Figure 13.3 for a graphical depiction of the relationship between odds ratio and absolute risk difference.

19.6

Assessing Absolute Diagnostic Yield: Cohort Study

- Patient $i = 1, 2, 3, \dots, n$
- In-sample sufficient statistics: $\text{pre}(X_1), \dots, \text{pre}(X_n), \text{post}(X_1, T_1), \dots, \text{post}(X_n, T_n)$

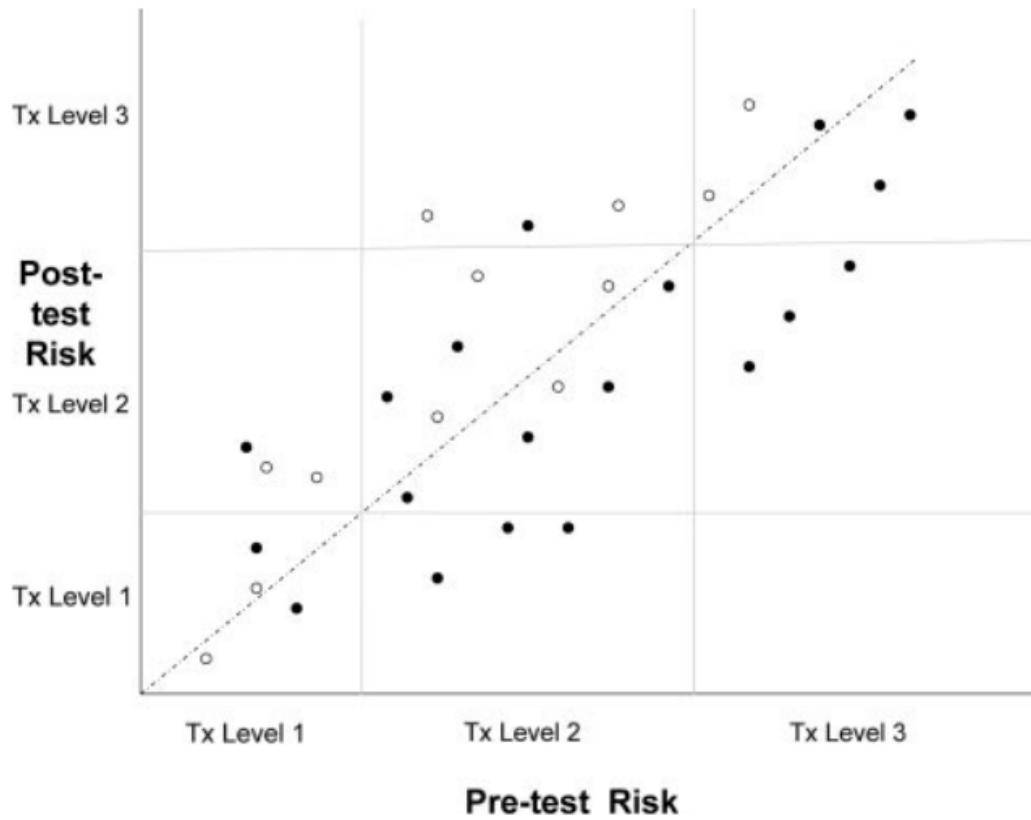
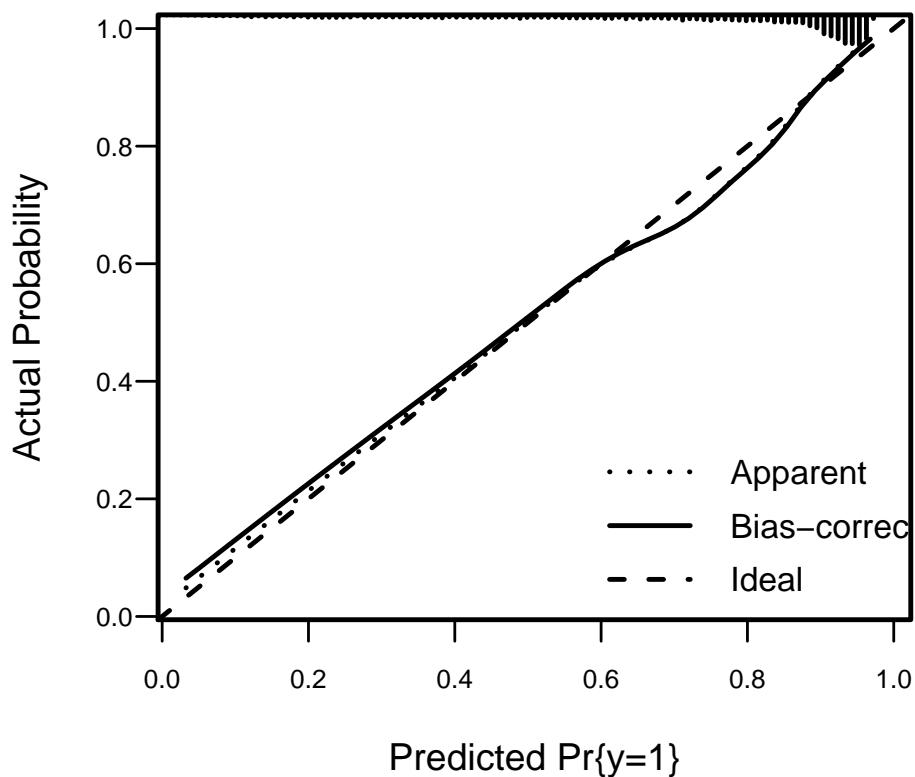


Figure 19.4: Pre vs. post-test probability. This may be summarized with quantile regression to estimate 0.1 and 0.9 quantiles of `post` as a function of `pre`. From Hlatky et al. [44]

Assessments assume that risk estimates are well calibrated, using, for example a high-resolution continuous calibration curve.



Out-of-sample assessment: compute $\text{pre}(X)$ and $\text{post}(X, T)$ for any X and T of interest

Summary measures

- quantile regression (55) curves as a function of pre
- overall mean $|\text{post} - \text{pre}|$
- quantiles of $\text{post} - \text{pre}$
- du_{50} : **distribution** of post when $\text{pre} = 0.5$
diagnostic utility at maximum pre-test uncertainty
 - Choose X so that $\text{pre} = 0.5$
 - Examine distribution of post at this pre
 - Summarize with quantiles, Gini's mean difference on prob. scale
 - Special case where test is binary (atypical): compute post for T^+ and for T^-

19.7

Assessing Diagnostic Yield: Case-Control & Other Oversampling Designs

- Intercept α is meaningless
- Choose X and solve for α so that $\text{pre} = 0.5$
- Proceed as above to estimate du_{50}

19.8

Example: Diagnosis of Coronary Artery Disease (CAD): Test = Total Cholesterol

```
require(rms)

getHdata(acath)
acath <- subset(acath, !is.na(choleste))
dd <- datadist(acath); options(datadist='dd')
f <- lrm(sigdz ~ rcs(age,5)*sex, data=acath)
pre <- predict(f, type='fitted')
g <- lrm(sigdz ~ rcs(age,4)*sex + rcs(choleste,4) + rcs(age,4) %ia%
           rcs(choleste,4), data=acath)
ageg <- c(40, 70)
psig <- Predict(g, choleste, age=ageg)
s <- lrm(tvdlm ~ rcs(age,4)*sex + rcs(choleste,4) + rcs(age,4) %ia%
           rcs(choleste,4), data=acath)
psev <- Predict(s, choleste, age=ageg)
ggplot(rbind('Significant CAD'=psig, '3 Vessel or Left Main CAD'=psev),
       adj.subtitle=FALSE)
```

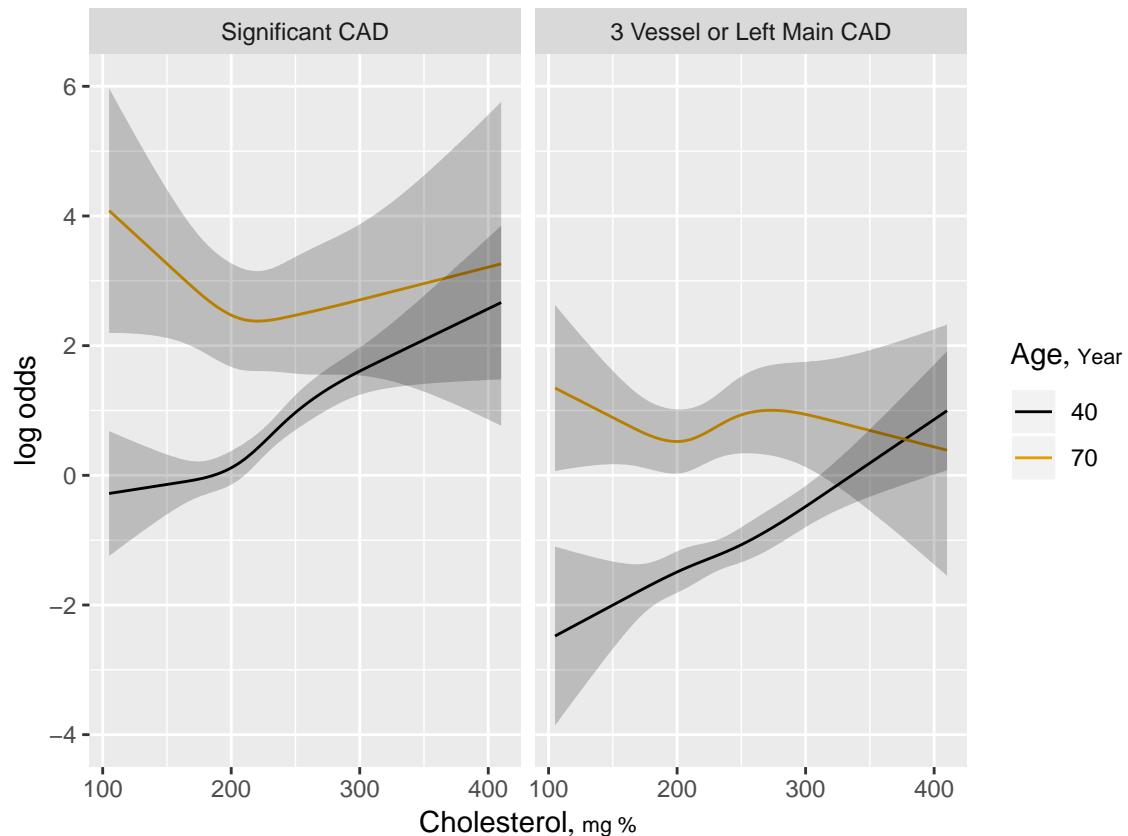


Figure 19.5: Relative effect of total cholesterol for age 40 and 70; Data from Duke Cardiovascular Disease Databank, $n = 2258$

```

post ← predict(g, type='fitted')
plot(pre, post, xlab='Pre-Test Probability (age + sex)',
      ylab='Post-Test Probability (age + sex + cholesterol)', pch=46)
abline(a=0, b=1, col=gray(.8))
lo ← Rq(post ~ rcs(pre, 7), tau=0.1) # 0.1 quantile
hi ← Rq(post ~ rcs(pre, 7), tau=0.9) # 0.9 quantile
at ← seq(0, 1, length=200)
lines(at, Predict(lo, pre=at)$yhat, col='red', lwd=1.5)
lines(at, Predict(hi, pre=at)$yhat, col='red', lwd=1.5)
abline(v=.5, col='red')

```

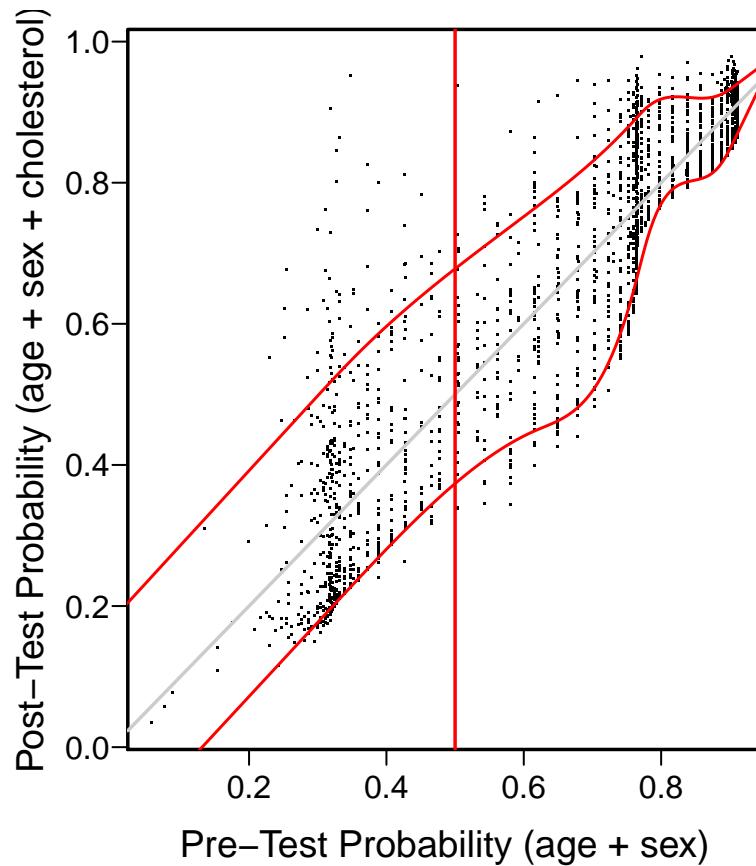


Figure 19.6: Diagnostic Utility of Cholesterol for Diagnosing Significant CAD. Curves are 0.1 and 0.9 quantiles from quantile regression using restricted cubic splines

19.9

Summary

Diagnostic utility needs to be estimated using measures of relevance to individual decision makers. Improper accuracy scoring rules lead to suboptimal decisions. Traditional risk modeling is a powerful tool in this setting. Cohort studies are ideal but useful measures can be obtained even with oversampling. Avoid categorization of any continuous or ordinal variables.

Chapter 20

Challenges of Analyzing High-Dimensional Data

Biomarker Uncertainty Principle:

A molecular signature can be either parsimonious or predictive, but not both.

[FE Harrell, 2009](#)

We have more data than ever, more good data than ever, a lower proportion of data that are good, a lack of strategic thinking about what data are needed to answer questions of interest, sub-optimal analysis of data, and an occasional tendency to do research that should not be done.

[FE Harrell, 2015](#)

20.1

Background

High-dimensional data are of at least three major types:

- Data collected on hundreds of thousands or millions of subjects with a diverse array of variables
- Time series of biologic signals collected every few milliseconds
- Extremely large data arrays where almost all the variables are of one type (the main topic here)

The last data class includes such areas as

- functional imaging
- gene microarray
- SNPs for genome-wide association studies
- RNA seq
- exome sequencing
- mass spectrometry

The research yield of analysis of such data has been disappointing to date, for many reasons such as:

- Biology is complex
- Use of non-reproducible research methodology
- Use of unproven statistical methods

- Multiple comparison problems and double dipping^a
- Failure to demonstrate value of information over and above the information provided by routine clinical variables, blood chemistry, and other standard medical tests
- Inadequate sample size for the complexity of the analytic task
- Overstatement of results due to searching for “winning” markers without understanding bias, regression to the mean (Section 14.8), and overfitting

Regarding double dipping/multiplicity a beautiful example is the dead salmon fMRI study by Bennett *et al.* in which the standard method of analyzing fMRI data by voxels was shown to find brain activation regions even in the dead brain (see <http://prefrontal.org/files/posters/Bennett-Salmon-2009.pdf>).

^aDouble dipping refers to using the same data to test a hypothesis that was formulated from a confession from the tortured dataset.

SCIENCE

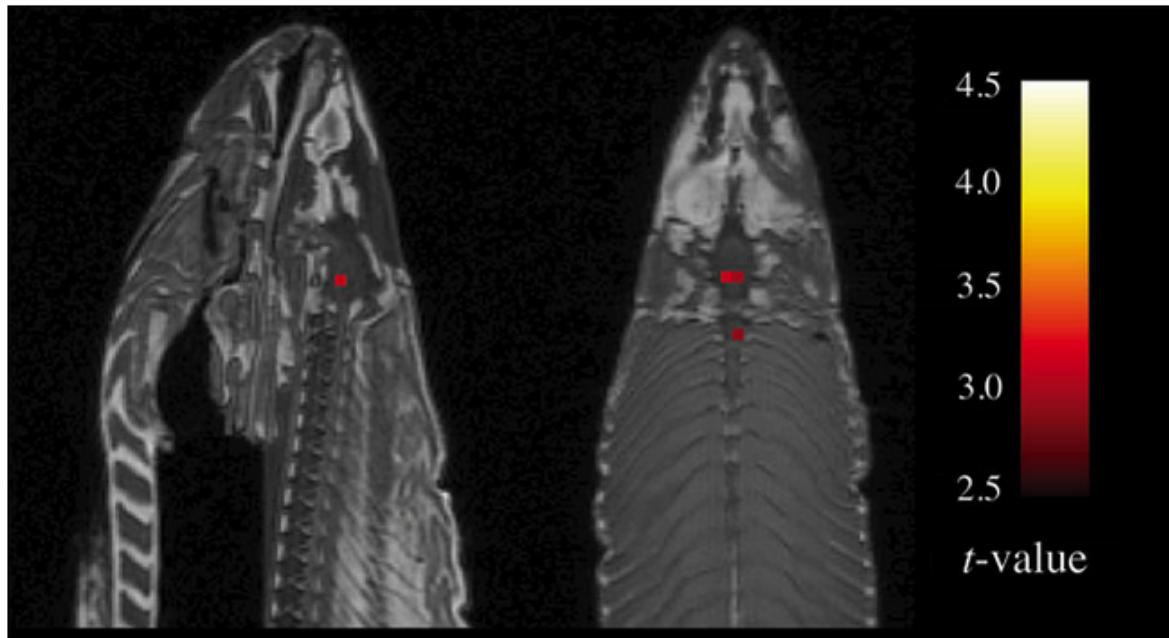
Brains and Behavior

fish

fMRI

Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings

BY ALEXIS MADRIGAL 09.18.09 | 5:37 PM | PERMALINK



Neuroscientist Craig Bennett purchased a whole Atlantic salmon, took it to a lab at Dartmouth, and put it into an fMRI machine used to study the brain. The beautiful fish was to be the lab's test object as they worked out some new methods.

So, as the fish sat in the scanner, they showed it "a series of photographs depicting human individuals in social situations." To maintain the rigor of the protocol (and

Wired, 2009-09-18.

Unfortunately, a large number of papers in genomics and imaging research have pretended that feature selection has no randomness, and have validated predictions in the same dataset used to find the signals, without informing the bootstrap or cross-validation procedure about the data mining used to derive the model so that the resampling procedure could repeat all feature selection steps afresh for each resample⁶. Only by re-running all data analysis steps that utilized Y for each resample can an unbiased estimate of future model performance be obtained.

20.2

Global Analytic Approaches

Let X_1, X_2, \dots, X_p denote an array of candidate features (gene expressions, SNPs, brain activation, etc.) and Y the response, diagnosis, or patient outcome being predicted.

20.2.1

One-at-a-Time Feature Screening

OaaT feature screening involves choosing a measure of association such as Pearson's χ^2 statistic^b, computing all p association measures against Y , and choosing those X features whose association measure exceeds a threshold. This is by far the most popular approach to feature discovery (and often to prediction, unfortunately) in genomics and functional imaging. It is demonstrably the worst approach in terms of the reliability of the "winning" and "losing" feature selection and because it results in poor predictive ability. The problems are due to multiple comparison problems, bias, typically high false negative rates, and to the fact that most features "travel in packs", i.e., act in networks rather than individually. As if OaaT could not get any worse, many practitioners create "risk scores" by simply adding together individual regression coefficients computed from individual "winning" X features without regard to the correlation structure of the X s. This gives far too much weight to the selected X s that are co-linear with each other.

There is a false belief that preventing a high false discovery rate solves the problems of OaaT. Most researchers fail to consider, among other things, the high false negative rate caused by their design and sample size.

OaaT results in highly overestimated effect sizes for winners, due to double dipping.

Multiplicity Corrections

In the belief that false positive discoveries are less desirable than missed discoveries, researchers employ multiplicity corrections to P -values arising from testing a large number of associations with Y . The most conservative approach uses the addition or Bonfer-

^bOddly, many practitioners of OaaT choose the more conservative and much slower to compute Fisher's exact test.

roni inequality to control the *family-wise error risk* which is the probability of getting *any* false positive association when the global null hypothesis that there are no true associations with Y holds. This can be carried out by testing the p associations against $\frac{\alpha}{p}$ where $\alpha = 0.05$ for example. A less conservative approach uses the *false discovery rate* (FDR), better labeled the *false discovery risk* or *false discovery probability*. Here one controls the fraction of false positive associations. For example one sets the critical value for association statistics to a less severe threshold than dictated by Bonferroni's inequality, by allowing $\frac{1}{10}^{\text{th}}$ of the claimed associations (those with the smallest P -values) to be false positives.

It is an irony that the attempt to control for multiplicity has led not only to missed discoveries and abandonment of whole research areas but has resulted in increased bias in the “discovered” features’ effect estimates. When using stepwise regression, the bias in regression coefficients comes not from multiplicity problems arising when experimentwise $\alpha = 0.05$ is used throughout but from using an α cutoff $< \frac{1}{2}$ for selecting variables. By selecting variables on the basis of small P -values, many of the selected variables were selected because their effects were overestimated, then regression to the mean sets in.

20.2.2

Forward Stepwise Variable Selection

Forward stepwise variable selection that does an initial screening of all the features, adds the most significant one, then attempts to add the next most significant feature, after adjusting for the first one, and so on. This approach is unreliable. It is only better than OaaT in two ways: (1) a feature that did not meet the threshold for the association with Y without adjusting for other features may become stronger after the selection of other features at earlier steps, and (2) the resulting risk score accounts for co-linearities of selected features. On the downside, co-linearities make feature selection almost randomly choose features from correlated sets of X s, and tiny changes in the dataset will result in different selected features.

20.2.3

Ranking and Selection

Feature discovery is really a ranking and selection problem. But ranking and selection methods are almost never used. An example bootstrap analysis on simulated data is

presented later. This involves sampling with replacement from the combined (X, Y) dataset and recomputing for each bootstrap repetition the p association measures for the p candidate X s. The p association measures are ranked. The ranking of each feature across bootstrap resamples is tracked, and a 0.95 confidence interval for the rank is derived by computing the 0.025 and 0.975 quantiles of the estimated ranks.

This approach puts OaaT in an honest context that fully admits the true difficulty of the task, including the high cost of the false negative rate (low power). Suppose that features are ranked so that a low ranking means weak estimated association and a high ranking means strong association. If one were to consider features to have “passed the screen” if their lower 0.95 confidence limit for their rank exceeds a threshold, and only dismisses features from further consideration if the upper confidence limit for their rank falls below a threshold, there would correctly be a large middle ground where features are not declared either winners or losers, and the researcher would be able to only make conclusions that are supported by the data. Readers of the class *Absence of Evidence* paper⁴ will recognize that this ranking approach solves the problem.

An example of bootstrap confidence intervals where multivariable modeling was used rather than OaaT association measures is in the course notes for *Regression Modeling Strategies*, Section 5.4. An example where the bootstrap is used in the context of OaaT is below.

20.2.4

Joint modeling of All Features Simultaneously using Shrinkage

This approach uses multivariable regression models along with penalized maximum likelihood estimation, random effects / hierarchical modeling, or skeptical Bayesian prior distributions for adjusted effects of all p X features simultaneously. Feature effects (e.g., log odds ratios) are discounted so as to prevent overfitting/over-interpretation and to allow these effects to be trusted out of context. The result is a high-dimensional regression model that is likely to provide well-calibrated absolute risk estimates and near-optimum predictive ability. Some of the penalization methods used are

1. lasso: a penalty on the absolute value of regression coefficient that highly favors zero as an estimate. This results in a large number of estimated coefficients being exactly zero, i.e., results in feature selection. The resulting parsimony may be

illusory: bootstrap repetition may expose the list of “selected” features to be highly unstable.

2. ridge regression (penalty function that is quadratic in the regression coefficients): does not result in a parsimoneous model but is likely to have the highest predictive value
3. elastic net: a combination of lasso and quadratic penalty that has some parsimony but has better predictive ability than the lasso. The difficulty is simultaneously choosing two penalty parameters (one for absolute value of β s, one for their sum of squares).

20.2.5

Random Forest

This is an approach that solves some of the incredible instability and low predictive power of individual regression trees. The basic idea of random forest is that one fits a regression tree using recursive partitioning (CART) on multiple random samples of **candidate features**. Multiple trees are combined. The result is no longer a tree; it is an uninterpretable black box. But in a way it automatically incorporates shrinkage and is often competitive with other methods in predictive ability.

There is evidence that minimal-assumption methods such as random forests are “data hungry”, requiring as many as 200 events per candidate variable for their apparent predictive discrimination to not decline when evaluated in a new sample¹⁰⁶.

20.2.6

Data Reduction Followed by Traditional Regression Modeling

This approach uses techniques such as principle component analysis (PCA) whereby a large number of candidate X s are reduced to a few summary scores. PCA is based on additively combining features so as to maximize the variation of the whole set of features that is explained by the summary score. A small number of PCA scores are then put into an ordinary regression model (e.g., binary logistic model) to predict Y . The result is sometimes satisfactory though no easier to interpret than shrinkage methods.

20.2.7

Model Approximation

Also called *pre-conditioning*, this method is general-purpose and promising^{79,96,5}. One takes a well-performing black box (e.g., random forest or full penalized regression with p features) that generates predicted responses \hat{Y} and incorporates the right amount of shrinkage to make the predictions well-calibrated. Then try to find a smaller set of X s that can represent \hat{Y} with high accuracy (e.g., $R^2 \geq 0.9$). Forward stepwise variable selection may be used for this purpose^c. This sub-model is an approximation to the “gold-standard” full black box. The ability to find a well-performing approximation is a test of whether the predictive signal is parsimoneous. If one requires 500 X s to achieve an $R^2 \geq 0.9$ in predicting the gold-standard predictions \hat{Y} , then it is not possible to be parsimoneous and predictive.

A major advantage of model approximation is that if the original complex model was well calibrated by using appropriate shrinkage, the smaller approximate model inherits that shrinkage.

20.2.8

Incorporating Biology into High-Dimensional Regression Models

This approach is likely to result in the most trustworthy discoveries as well as the best predictive accuracy, if existing biological knowledge is adequate for specification of model structure. This is a structured shrinkage method where pathway (e.g., gene pathway) information is inserted directly in the model. One may encode multiple paths into a simultaneous regression model such that genes are “connected” to other genes in the same pathway. This allows an entire path to be emphasized or de-emphasized.

^cHere one is developing a mechanistic prediction where the true R^2 is 1.0.

20.3

Simulated Examples

Monte Carlo simulation, when done in a realistic fashion, has the great advantage that one knows the truth, i.e., the true model and values of model parameters from which the artificial population was simulated. Then any derived model or estimates can be compared to the known truth. Also with simulation, one can easily change the sample size being simulated so that the effect of sample size can be studied and an adequate sample size that leads to reliable results can be computed. One can also easily change the dimensionality of the features.

20.3.1

Simulation To Understand Needed Sample Sizes

One of the most common association measures used in genomic studies is the odds ratio. As shown in Section 6.8 and Figure 6.1 , the odds ratio (OR) is very difficult to estimate when the outcome is rare or when a binary predictive feature has a prevalence far from $\frac{1}{2}$. That is for the case when only a single pre-specified is estimated. When screening multiple features for interesting associations, one is effectively estimating a large number of ORs, and in order to make correct decisions about which features are promising and which aren't, one must be able to control the margins of error of the entire set of OR estimates.

In the following simulation consider varying sample size n and number of candidate features p . We simulate p binary features with known true ORs against the diagnosis or outcome Y . The true unknown ORs are assumed to have a $\text{normal}(\mu = 0, \sigma = 0.25)$ distribution. We want to judge the ability to jointly estimate p associations and to rank order features by observed associations. The analysis that is simulated does not examine multiple X s simultaneously, so we save time by simulating just the total numbers of zeros and ones for each X , given Y .

```
# For a vector of n binary outcomes y, simulates p binary features
# x that have a p-vector of fixed prevalences l | y=0 of prev and are connected
# to y by a p-vector of true population odds ratios ors.
# Estimates the p odds ratios against the simulated outcomes and
# returns a data frame summarizing the information
#
# Note: the odds ratio for x predicting y is the same as the odds ratio
# for y predicting x. y is simulated first so that all features will
```

```
# be analyzed against the same outcomes

sim <- function(y, prev, or) {
  n <- length(y)
  p <- length(prev)
  if(p != length(or)) stop('prev and or must have the same length')

  # prev = Pr(x=1 | y=0); let the odds for this be oprev = prev / (1-prev)
  # or = odds(x=1 | y=1) / oprev
  # Pr(x=1 | y=1) = oprev / ((1 / or) + oprev)

  oprev <- prev / (1 - prev)
  p1 <- oprev / ((1 / or) + oprev)
  n0 <- sum(y == 0)
  n1 <- sum(y == 1)
  # For n0 observations sample x so that Pr(x=0 | y=0) = prev
  nxy0 <- rbinom(p, size=n0, prob=prev)
  nxy1 <- rbinom(p, size=n1, prob=p1)

  # Compute p sample odds ratios
  sor <- (n0 - nxy0) * nxy1 / (nxy0 * (n1 - nxy1))
  g <- function(x) ifelse(x ≥ 1, x, 1 / x)
  r1 <- rank(sor)[which.max(or)] / p
  r2 <- rank(or)[which.max(sor)] / p
  data.frame(prev, or, nx=nxy0 / n0, obsprev0=(nxy0 + nxy1) / n,
             obsprev=nxy1 / (nxy0 + nxy1), obsor=sor, n=n,
             N = paste('n', n, sep=':'),
             Features=paste('Features', p, sep=':'),
             mmoe = quantile(g(sor / or), 0.90, na.rm=TRUE),
             obsranktrue=r1, truerankobs=r2,
             rho=cor(sor, or, method='spearman', use='pair'))
}

}
```

```
U <- NULL
set.seed(1)
for(n in c(50, 100, 250, 500, 1000, 2000)) {
  for(p in c(10, 50, 500, 1000, 2000)) {
    for(yprev in c(.1, .3)) {
      y <- rbinom(n, 1, yprev)
      prev <- runif(p, .05, .5)
      or <- exp(rnorm(p, 0, .25))
      u <- cbind(sim(y, prev, or),
                 Yprev=paste('Prevalence of Outcome', yprev, sep=':'))
      U <- rbind(U, u)
    }
  }
}
```

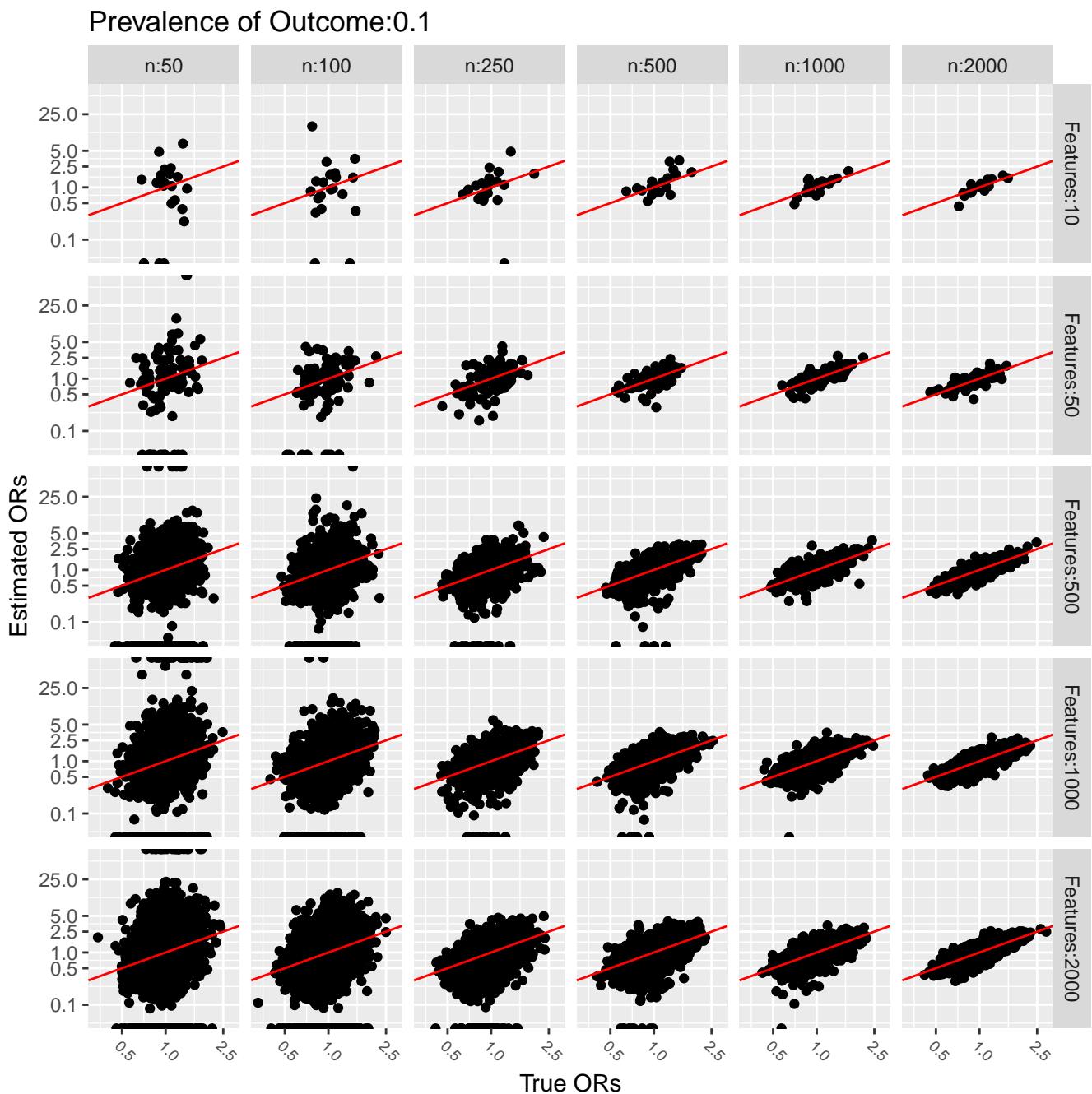
In the plots below red lines show the line of identity.

```
require(ggplot2)
pl <- function(yprev) {
  br <- c(.01, .1, .5, 1, 2.5, 5, 25, 100)
  ggplot(subset(U, Yprev=yprev),
         aes(x=or, y=obsor)) + geom_point() + facet_grid(Features ~ N) +
         ggttitle(paste('Prevalence of Outcome', yprev, sep=':')) +
```

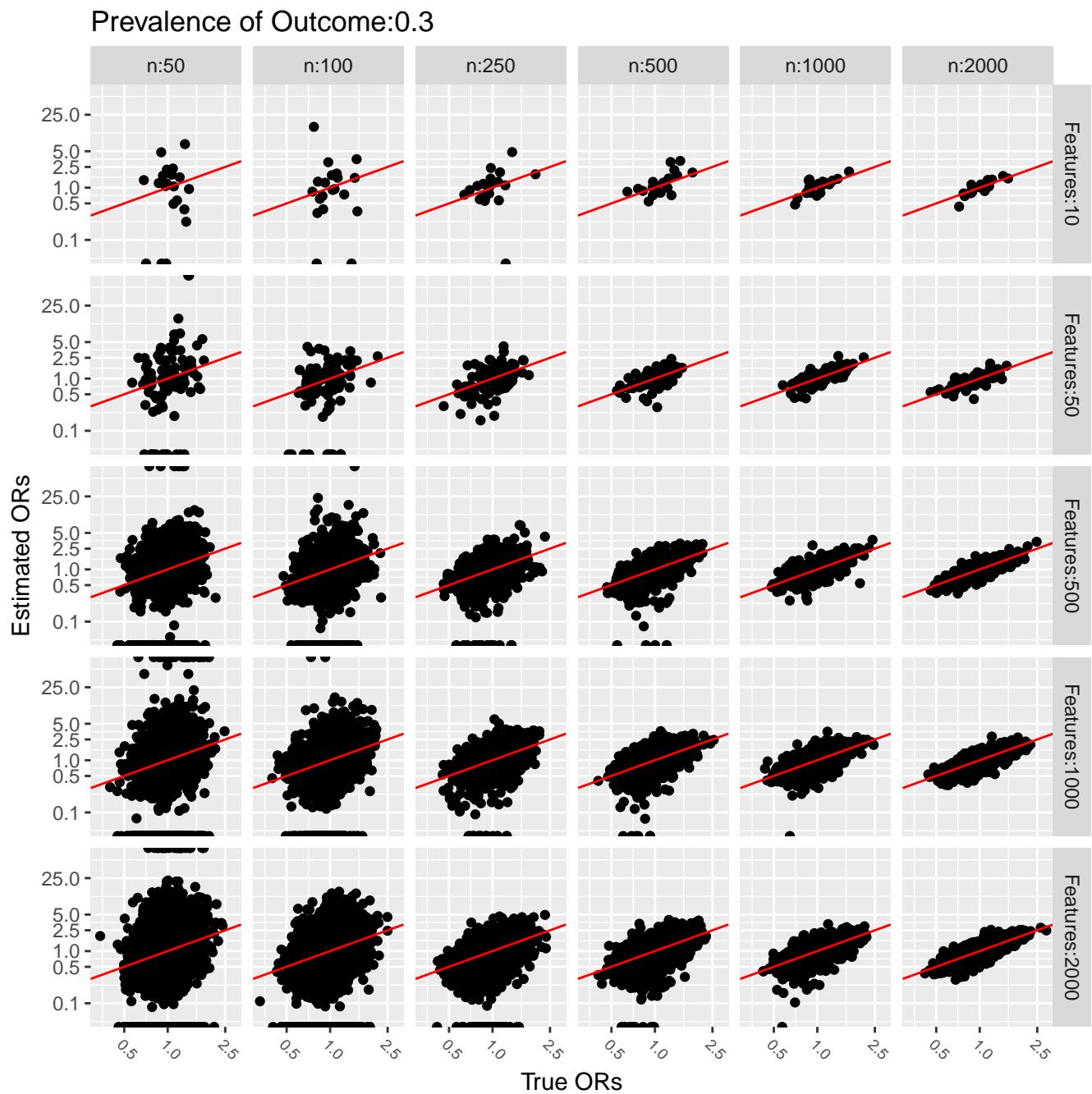
```

xlab('True ORs') + ylab('Estimated ORs') +
scale_x_log10(breaks=br) + scale_y_log10(breaks=br) +
theme(axis.text.x = element_text(size = rel(0.8), angle=-45,
hjust=0, vjust=1)) +
geom_abline(col='red')
}
p1(0.1)

```



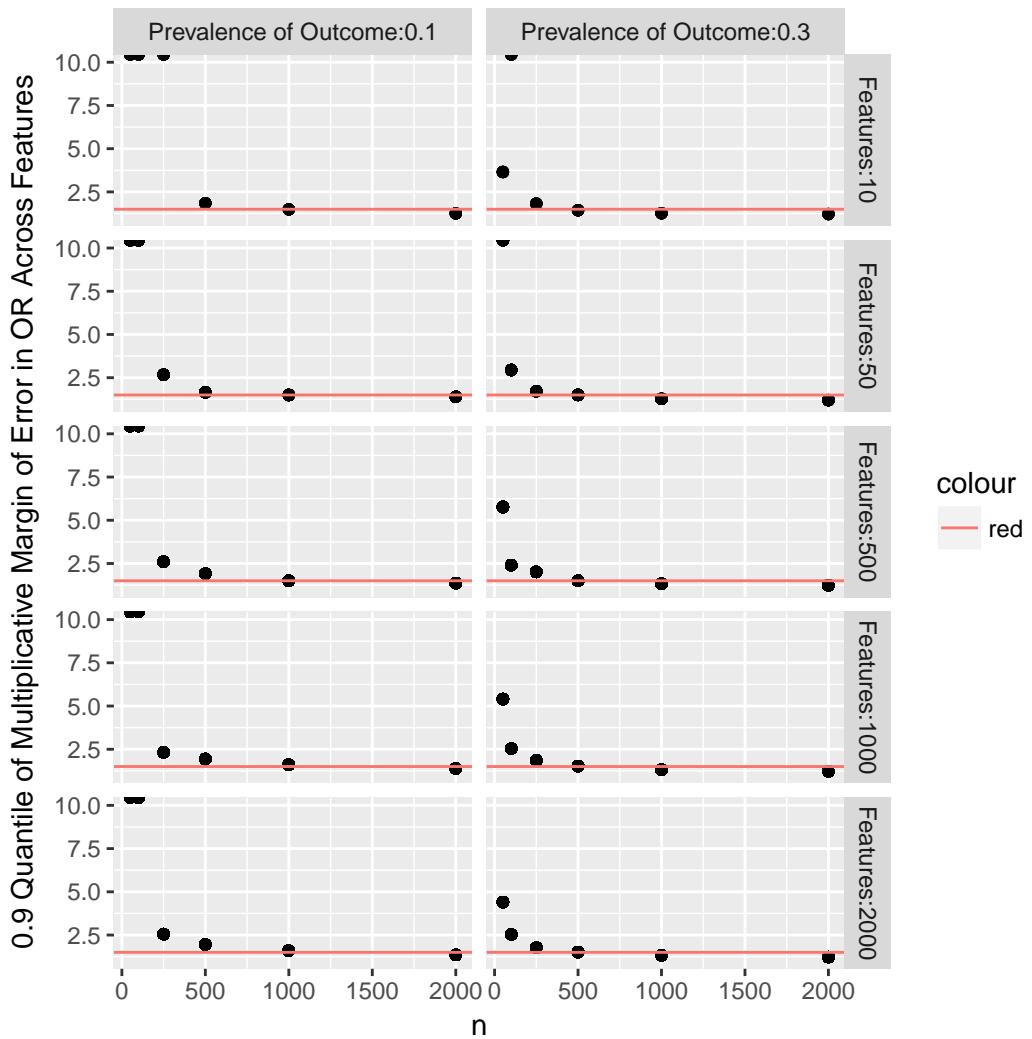
```
p1(0.3)
```



The last two figures use a log scale for the y -axis (estimated odds ratios), so the errors in estimating the odds ratios are quite severe. For a sample size of $n = 50$ one cannot even estimate a single pre-specified odds ratio. To be able to accurately assess 10 ORs (10 candidate features) requires about $n = 1000$. To assess 2000 features, a sample size of $n = 2000$ seems adequate only for the very smallest and very largest true ORs.

The plot below summarizes the previous plots by computing the 0.9 quantile of the multiplicative margin of error (fold change) over the whole set of estimated odds ratios, ignoring direction.

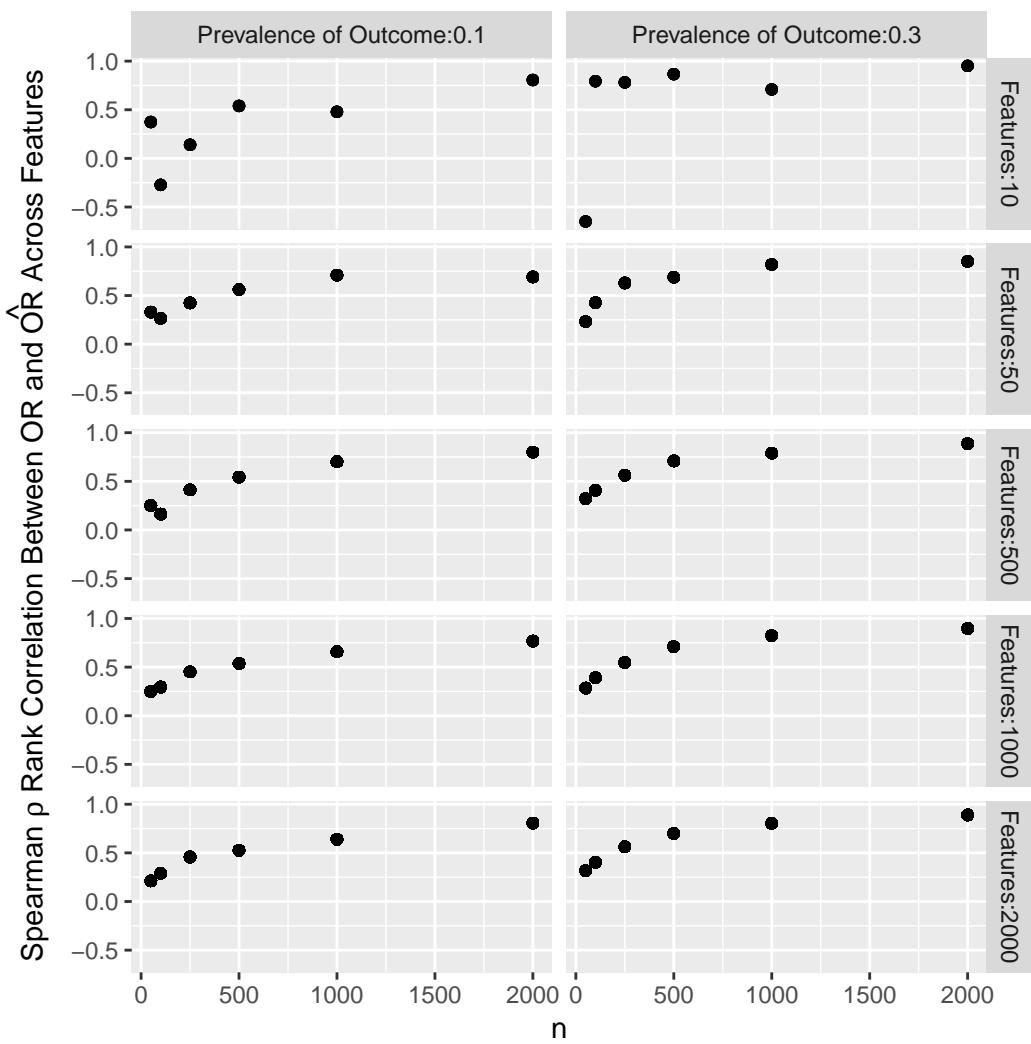
```
ggplot(U, aes(x=n, y=mmoe)) + geom_point() + facet_grid(Features ~ Yprev) +
  geom_hline(aes(yintercept=1.5, col='red')) +
  ylim(1, 10) +
  ylab('0.9 Quantile of Multiplicative Margin of Error in OR Across Features')
```



Horizontal red lines depict a multiplicative margin of error (MMOE) of 1.5 which may be considered the minimally acceptable error in estimating odds ratios. This was largely achieved with $n = 1000$ for a low-incidence Y , and $n = 500$ for a moderate-incidence Y .

Another way to summarize the results is to compute the Spearman rank correlation between estimated and true underlying odds ratios over the entire set of estimates.

```
ggplot(U, aes(x=n, y=rho)) + geom_point() +
  facet_grid(Features ~ Yprev) +
  ylab(expression(paste('Spearman ', rho, ' Rank Correlation Between ',
    'OR, ', ' and ', 'hat(OR), ' Across Features')))
```



One may desire a correlation with the truth of say 0.8 and can solve for the needed sample size.

20.3.2

Bootstrap Analysis for One Simulated Dataset

Suppose that one wanted to test p candidate features and select the most “significant” one for a validation study. How likely is the apparently “best” one to be truly the best? What is a confidence interval for the rank of this “winner”? How much bias in the OR does the selection process create? The bootstrap can be used to answer all of these questions without needing to assume anything about true population parameter values. The bootstrap can take into account many sources of uncertainty. We use the bootstrap to estimate the bias in the apparent highest and apparent lowest odds ratios—the two “winners”. The sample size of the simulated data is 600 subjects and

there are 300 candidate features.

The bootstrap is based on sampling with replacement from the rows of the entire data matrix (X, Y). In order to sample from the rows we need to generate raw data, not just numbers of “successes” and “failures” as in the last simulation.

```
# Function to simulate the raw data
# prev is the vector of prevalences of x when y=0 as before
# yprev is the overall prevalence of y
# n is the sample size
# or is the vector of true odds ratios
sim <- function(n, yprev, prev, or) {
  y <- rbinom(n, 1, yprev)
  p <- length(prev)
  if(p != length(or)) stop('prev and or must have the same length')

  # prev = Pr(x=1 | y=0); let the odds for this be oprev = prev / (1-prev)
  # or = odds(x=1 | y=1) / oprev
  # Pr(x=1 | y=1) = oprev / ((1 / or) + oprev)

  oprev <- prev / (1 - prev)
  p1 <- oprev / ((1 / or) + oprev)
  x <- matrix(NA, nrow=n, ncol=p)
  for(j in 1 : p)
    x[, j] <- ifelse(y == 1, rbinom(n, 1, prob = p1[j]),
                      rbinom(n, 1, prob = prev[j]))
  list(x=x, y=y)
}

# Function to compute the sample odds ratios given x matrix and y vector
ors <- function(x, y) {
  p <- ncol(x)
  or <- numeric(p)
  for(j in 1 : p) {
    f <- table(x[, j], y)
    or[j] <- f[2, 2] * f[1, 1] / (f[1, 2] * f[2, 1])
  }
  or
}

# Generate sample of size 600 with 300 features
# Log odds ratios have a normal distribution with mean 0 SD 0.3
# x have a random prevalence uniform [0.05, 0.5]
# y has prevalence 0.3

set.seed(188)
n <- 600; p <- 300
prev <- runif(p, .05, .5)
or <- exp(rnorm(p, 0, .3))
z <- sim(n, 0.3, prev, or)

# Compute estimated ORs
x <- z$x; y <- z$y
sor <- ors(x, y)
# Show how estimates related to true ORs
ggplot(data.frame(or, sor), aes(x=or, y=sor)) + geom_point() +
```

```

xlab('True OR') + ylab('Estimated OR')

# Print the largest estimated OR and its column number,
# and corresponding true OR, and similarly for the smallest.
largest   ← max(sor)
imax      ← which.max(sor)
true.imax ← or[imax]
mmoe.imax ← largest / true.imax
smallest  ← min(sor)
imin      ← which.min(sor)
true.imin ← or[imin]
mmoe.imin ← smallest / true.imin
cat('\nLargest observed OR\n')

```

Largest observed OR

```

cat('OR:', round(largest, 2), ' Feature #', imax, ' True OR:',
    round(true.imax, 2), ' MMOE:', round(mmoe.imax, 2), '\n')

```

OR: 2.94 Feature # 90 True OR: 1.71 MMOE: 1.72

```

cat('Rank of winning feature among true ORs:', sum(or ≤ or[imax]), '\n\n')

```

Rank of winning feature among true ORs: 285

```

cat('Smallest observed OR\n')

```

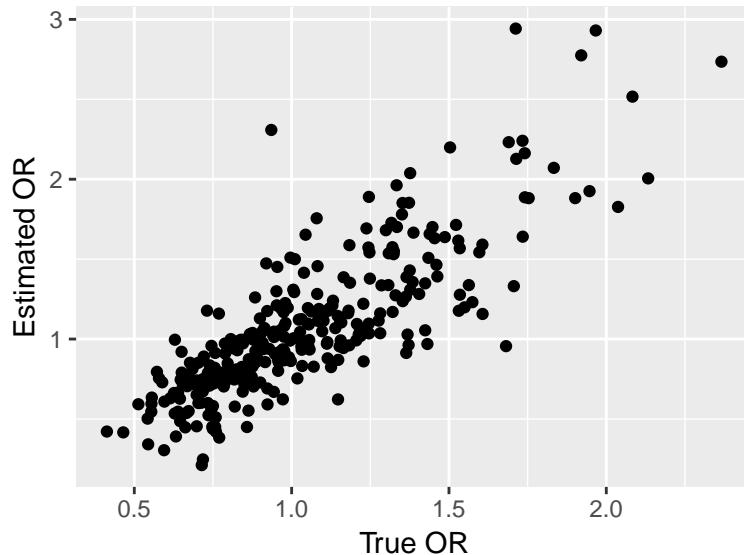
Smallest observed OR

```

cat('OR:', round(smallest, 2), ' Feature #', imin, ' True OR:',
    round(true.imin, 2), ' MMOE:', round(mmoe.imin, 2), '\n')

```

OR: 0.21 Feature # 99 True OR: 0.71 MMOE: 0.3



Next use the bootstrap to get an estimate of the MMOE for the observed largest OR, and a 0.95 confidence interval for the true unknown rank of the largest observed OR

from among all the features. 1000 bootstrap resamples are drawn. In estimating the MMOE we are estimating bias in the largest log odds ratio when a new “largest” OR is found in each bootstrap resample. That estimated OR is compared to the OR evaluated in the whole sample for the same column number. This is also done for the “smallest” OR.

```
set.seed(11)
B ← 1000
ranksS ← ranksL ← mmoeS ← mmoeL ← numeric(B)

for(k in 1 : B) {
  # Draw a sample of size n with replacement
  i ← sample(1 : n, n, replace=TRUE)
  # Compute sample ORs on the new sample
  bor ← ors(x[i, ], y[i])
  blargest ← max(bor)
  bmax ← which.max(bor)
  ranksL[k] ← sum(bor ≤ largest)
  mmoeL[k] ← blargest / sor[bmax]
  bsmallest ← min(bor)
  bmin ← which.min(bor)
  ranksS[k] ← sum(bor ≤ smallest)
  mmoeS[k] ← bsmallest / sor[bmin]
}
```

The bootstrap geometric mean MMOE for the smallest odds ratio was zero due to small frequencies in some X s. The median bootstrap MMOE was used to bias-correct the observed smallest OR, while the geometric mean was used for the largest.

```
pr ← function(which, ranks, mmoe, mmoe.true, estor, or.true) {
  gm ← exp(mean(log(mmoe)))
  cat(which, 'OR\n')
  cat('CL for rank:', quantile(ranks, c(0.025, 0.975)),
      ' Median MMOE:', round(median(mmoe), 2),
      ' Geometric mean MMOE:', round(gm, 2),
      '\nTrue MMOE:', round(mmoe.true, 2), '\n')
  bmmoe ← if(which == 'Largest') gm else median(mmoe)
  cat('Bootstrap bias-corrected', tolower(which), 'OR:',
      round(estor / bmmoe, 2),
      ' Original OR:', round(estor, 2),
      ' True OR:', round(or.true, 2),
      '\n\n')
}
pr('Largest', ranksL, mmoeL, mmoe.imax, largest, true.imax)
```

```
Largest OR
CL for rank: 294 299    Median MMOE: 1.45    Geometric mean MMOE: 1.52
True MMOE: 1.72
Bootstrap bias-corrected largest OR: 1.94    Original OR: 2.94    True OR: 1.71
```

```
pr('Smallest', ranksS, mmoeS, mmoe.imin, smallest, true.imin)
```

```
Smallest OR
```

```
CL for rank: 0 5    Median MMOE: 0.32    Geometric mean MMOE: 0
True MMOE: 0.3
Bootstrap bias-corrected smallest OR: 0.67    Original OR: 0.21    True OR: 0.71
```

The bootstrap bias-corrected ORs took the observed extreme ORs and multiplied them by their respective bootstrap geometric mean or median MMOEs. The bias-corrected estimates are closer to the true ORs.

The data are consistent with the observed smallest OR truly being in the bottom 5 and the observed largest OR truly being in the top 7.

Here is some example wording that could be used in a statistical analysis plan: We did not evaluate the probability that non-selected genes are truly unimportant but will accompany the planned gene screening results with a bootstrap analysis to compute 0.95 confidence intervals for the rank of each gene, using a non-directional measure of association when ranking. The “winning” genes will have high ranks in competing with each other (by definition) but if the lower confidence limits include mid- to low-level ranks the uncertainty of these winners will be noted. Conversely, if the upper confidence intervals of ranks of “losers” extend well into the ranks of “winners”, the uncertainty associated with our ability to rule out “losers” for further investigation will be noted.

Chapter 21

Reproducible Research

Disconfirmation bias: giving expected results a relatively free pass but rigorously checking non-intuitive results

Nuzzo [75]

An excellent article on how to do reproducible research is⁷⁰ for which the pdf file is openly available. The link to an excellent video by Garrett Grolemund is on the right.



21.1

Non-reproducible Research

- Misunderstanding statistics
- “Investigator” moving the target
- Lack of a blinded analytic plan
- Tweaking instrumentation / removing “outliers”
- Floating definitions of response variables
- Pre-statistician “normalization” of data and background subtraction
- Poorly studied high-dimensional feature selection
- Programming errors
- Lack of documentation
- Failing to script multiple-step procedures
 - using spreadsheets and other interactive approaches for data manipulation
- Copying and pasting results into manuscripts
- Insufficient detail in scientific articles
- No audit trail

21.2

General Importance of Sound Methodology

21.2.1

Translation of Research Evidence from Animals to Humans

- Screened articles having preventive or therapeutic intervention in in vivo animal model, > 500 citations (Hackam and Redelmeier [39])
- 76 “positive” studies identified
- Median 14 years for potential translation
- 37 judged to have good methodological quality (flat over time)
- 28 of 76 replicated in human randomized trials; 34 remain untested
- ↑ 10% methodology score ↑ odds of replication $\times 1.28$ (0.95 CL 0.97–1.69)
- Dose-response demonstrations: ↑ odds $\times 3.3$ (1.1–10.1)

Note: The article misinterpreted P -values

21.2.2

Other Problems

- Rhine and ESP: “the student’s extra-sensory perception ability has gone through a marked decline”
- Floating definitions of X or Y : association between physical symmetry and mating behavior; acupuncture
- Selective reporting and publication bias

- Journals seek confirming rather than conflicting data
- Damage caused by hypothesis tests and cutoffs
- Ioannidis: $\frac{1}{3}$ of articles in *Nature* never get **cited**, let alone replicated
- Biologic and lab variability
- Unquestioning acceptance of research by the “famous”
 - Weak coupling ratio exhibited by decaying neutrons fell by 10 SDs from 1969–2001

21.2.3

What's Gone Wrong with Omics & Biomarkers?

- Gene expression-based prognostic signatures in lung cancer: Ready for clinical use? (Subramanian and Simon [102])
- NSCLC gene expression studies 2002–2009, $n \geq 50$
- 16 studies found
- Scored on appropriateness of protocol, stat validation, medical utility
- Average quality score: 3.1 of 7 points
- No study showed prediction improvement over known risk factors; many failed to validate
- Most studies did not even consider factors in guidelines
 - Completeness of resection only considered in 7
 - Similar for tumor size
 - Some only adjusted for age and sex

21.2.4

Failure of Replication in Preclinical Cancer Research

- Scientists at Amgen tried to confirm published findings related to a line of research, before launching development
- Identified 53 ‘landmark’ studies
- Scientific findings confirmed in only 6 studies
- Non-reproduced articles cited far more frequently than reproduced articles

Begley CG, Ellis LM: Raise standards for preclinical cancer research.

Nature 483:531-533; 2012

Natural History of New Fields

Each new field has a rapid exponential growth of its literature over 5–8 years (“new field phase”), followed by an “established field” phase when growth rates are more modest, and then an “over-maturity” phase, where the rates of growth are similar to the growth of the scientific literature at large or even smaller. There is a parallel in the spread of an infectious epidemic that emerges rapidly and gets established when a large number of scientists (and articles) are infected with these concepts. Then momentum decreases, although many scientists remain infected and continue to work on this field. New omics infections continuously arise in the scientific community.

Ionnidis [51]

The New York Times

April 16, 2012

A Sharp Rise in Retractions Prompts Calls for Reform

By CARL ZIMMER

In the fall of 2010, Dr. Ferric C. Fang made an unsettling discovery. Dr. Fang, who is editor in chief of the journal *Infection and Immunity*, found that one of his authors had doctored several papers.

It was a new experience for him. “Prior to that time,” he said in an interview, “*Infection and Immunity* had only retracted nine articles over a 40-year period.”

The journal wound up [retracting](#) six of the papers from the author, Naoki Mori of the University of the Ryukyus in Japan. And it soon became clear that *Infection and Immunity* was hardly the only victim of Dr. Mori’s misconduct. Since then, other scientific journals have retracted two dozen of his papers, [according to the watchdog blog Retraction Watch](#).

21.3

System Forces



21.4

Strong Inference

Cognitive biases are hitting the accelerator of science: the process spotting potentially important scientific relationships. Countering those biases comes down to strengthening the ‘brake’: the ability to slow down, be sceptical of findings and eliminate false positives and dead ends.

Nuzzo [75]

16 October 1964, Volume 146, Number 3642

SCIENCE

Strong Inference

Certain systematic methods of scientific thinking may produce much more rapid progress than others.

John R. Platt

“nature” or the experimental outcome chooses—to go to the right branch or the left; at the next fork, to go left or right; and so on. There are similar branch points in a “conditional computer program,” where the next move depends on the result of the last calculation. And there is a “conditional inductive tree” or “logical tree” of this kind written out in detail in many first-year chemistry books, in the table of steps for qualitative analysis of an unknown sample, where the student

Platt [81]

- Devise alternative hypotheses
- Devise an experiment with alternative possible outcomes each of which will exclude a hypothesis
- Carry out the experiment
- Repeat
- Regular, explicit use of alternative hypotheses & sharp exclusions → rapid & powerful progress
- “Our conclusions . . . might be invalid if . . . (i) . . . (ii) . . . (iii) . . . We shall describe experiments which eliminate these alternatives.”⁸¹

21.5

Pre-Specified Analytic Plans

I have enormous flexibility in how I analyze my data and what I choose to report. This creates a conflict of interest. The only way to avoid this is for me to tie my hands in advance. Precommitment to my analysis and reporting plan mitigates the influence of these cognitive biases.

Brian Nosek, Center for Open Science⁷⁵

- Long the norm in multi-center RCTs
- Needs to be so in **all** fields of research using data to draw inferences⁸⁶
- Front-load planning with investigator
 - too many temptations later once see results (e.g., $P = 0.0501$)
- SAP is signed, dated, filed
- Pre-specification of reasons for exceptions, with exceptions documented (when, why, what)
- Becoming a policy in VU Biostatistics

21.6

Summary

Methodologic experts have much to offer:

- Biostatisticians and clinical epidemiologists play important roles in
 - assessing the needed information content for a given problem complexity
 - minimizing bias
 - maximizing reproducibility
- For more information see:
 - ctspedia.org
 - reproducibleresearch.net
 - groups.google.com/group/reproducible-research

21.7

Software

21.7.1

Goals of Reproducible Analysis/Reporting

- Be able to reproduce your own results
- Allow others to reproduce your results

Time turns each one of us into another person, and by making effort to communicate with strangers, we help ourselves to communicate with our future selves.

Schwab and Claerbout

- Reproduce an entire report, manuscript, dissertation, book with a single system command when changes occur in:
 - operating system, stat software, graphics engines, source data, derived variables, analysis, interpretation
- Save time
- Provide the ultimate documentation of work done for a paper

See <http://biostat.mc.vanderbilt.edu/StatReport>

21.7.2

History of Literate Programming

- Donald Knuth found his own programming to be sub-optimal
- Reasons for programming attack not documented in code; code hard to read
- Invented **literate programming** in 1984
 - mix code with documentation in same file
 - “pretty printing” customized to each, using \TeX

- not covered here: a new way of programming
- Knuth invented the `noweb` system for combining two types of information in one file
 - *weaving* to separate non-program code
 - *tangling* to separate program code

See <http://www.ctan.org/tex-archive/help/LitProg-FAQ>

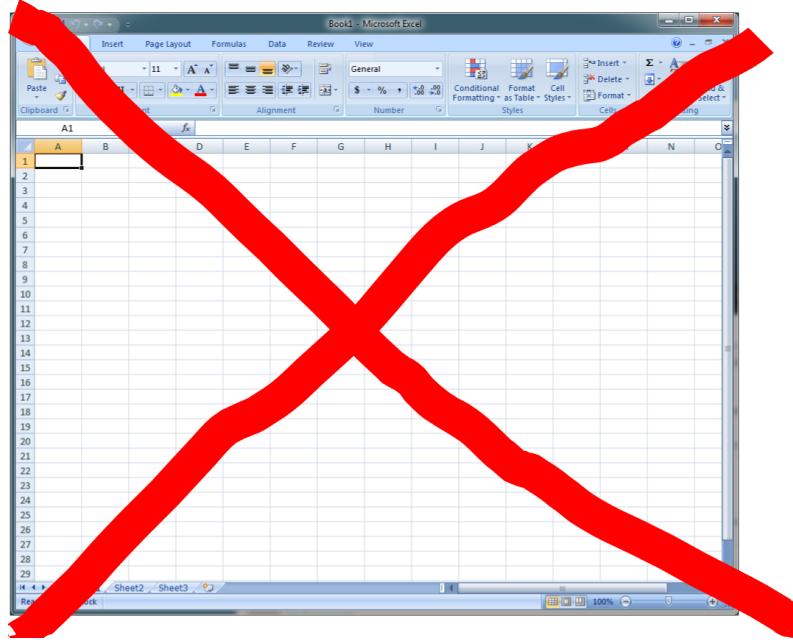
- Leslie Lamport made \TeX easier to use with a comprehensive macro package \LaTeX in 1986
- Allows the writer to concern herself with structures of ideas, not typesetting
- \LaTeX is easily modifiable by users: new macros, variables, *if-then* structures, executing system commands (Perl, etc.), drawing commands, etc.
- S system: Chambers, Becker, Wilks of Bell Labs, 1976
- R created by Ihaka and Gentleman in 1993, grew partly as a response to non-availability of S-Plus on Linux and Mac
- Friedrich Leisch developed Sweave in 2002
- Yihui Xie developed `knitr` in 2011

21.7.3

knitr Approach

- `knitr` is an R package on CRAN
- Uses `noweb` and an `sweave` style in \LaTeX ; see yihui.name/knitr,¹¹⁴ <http://yihui.github.com/knitr>
- `knitr` also works with Markdown and other languages
- `knitr` is tightly integrated into RStudio

A Bad Alternative to knitr



- *Insertions* are a major component
 - R printout after code chunk producing the output; plain tables
 - single pdf or postscript graphic after chunk, generates \LaTeX \includegraphics command
 - direct insertion of \LaTeX code produced by R functions
 - computed values inserted outside of code chunks
- Major advantages over Microsoft Word: composition time, batch mode, easily maintained scripts, beauty
- knitr produces self-documenting reports with nice graphics, to be given to clients
 - showing code demonstrates you are not doing “pushbutton” research

21.7.4

knitr Features

- R code set off by lines containing only <>>=

- \LaTeX text starts with a line containing only @
- knitr senses when a chunk produces a graphic (even without `print()`) and automatically includes the graphic in \LaTeX
- All other lines sent to \LaTeX verbatim, R code and output sent to \LaTeX by default but this can easily be overridden
- Can specify that a chunk produces markup that is directly typeset; this is how complex \LaTeX tables generated by R
- Can include calculated variables directly in sentences, e.g.
And the final answer is 3. will produce “And the final answer is 3.”
- Easy to customize chunk options and add advanced features such as automatically creating a \LaTeX figure environment if a caption is given in the chunk header
- Setup for advanced features, including code pretty-printing, shown at <http://biostat.mc.vanderbilt.edu/KnitrHowto>
- Simplified interface to `tikz` graphics
- Simplified implementation of caching
- More automatic pretty-printing; support for \LaTeX listings package built-in

See <http://biostat.mc.vanderbilt.edu/KnitrHowto>

Summary

Much of research that uses data analysis is not reproducible. This can be for a variety of reasons, the most major one being poor design and poor science. Other causes include tweaking of instrumentation, the use of poorly studied high-dimensional feature selection algorithms, programming errors, lack of adequate documentation of what was done, too much copy and paste of results into manuscripts, and the use of spreadsheets and other interactive data manipulation and analysis tools that do not provide a usable audit trail of how results were obtained. Even when a research journal allows the authors the “luxury” of having space to describe their methods, such text can never be specific enough for readers to exactly reproduce what was done. All too often, the authors themselves are not able to reproduce their own results. Being able to reproduce

an entire report or manuscript by issuing a single operating system command when any element of the data change, the statistical computing system is updated, graphics engines are improved, or the approach to analysis is improved, is also a major time saver.

It has been said that the analysis code provides the ultimate documentation of the “what, when, and how” for data analyses. Eminent computer scientist Donald Knuth invented literate programming in 1984 to provide programmers with the ability to mix code with documentation in the same file, with “pretty printing” customized to each. Lamport’s \LaTeX , an offshoot of Knuth’s \TeX typesetting system, became a prime tool for printing beautiful program documentation and manuals. When Friedrich Leisch developed Sweave in 2002, Knuth’s literate programming model exploded onto the statistical computing scene with a highly functional and easy to use coding standard using R and \LaTeX and for which the Emacs text editor has special dual editing modes using ESS. This approach has now been extended to other computing systems and to word processors. Using R with \LaTeX to construct reproducible statistical reports remains the most flexible approach and yields the most beautiful reports, while using only free software. One of the advantages of this platform is that there are many high-level R functions for producing \LaTeX markup code directly, and the output of these functions are easily directly to the \LaTeX output stream created by `knitr`.

21.8

Further Reading

An excellent book is Stodden, Leisch, and Peng [101]. See also

- <https://github.com/SISBID/Module3>: course by Baggerly and Broman
- reproducibleresearch.net
- cran.r-project.org/web/views/ReproducibleResearch.html
- www.nature.com/nature/focus/reproducibility
- biostat.mc.vanderbilt.edu/ReproducibleResearchTutorial
- biostat.mc.vanderbilt.edu/KnitrHowto
- biostat.mc.vanderbilt.edu/StatReport
- groups.google.com/forum/#!forum/reproducible-research
- resources.rstudio.com/rstudio-conf-2019/r-markdown-the-bigger-picture

Annotated Bibliography

- [1] C. J. Adcock. "Sample Size Determination: A Review". In: *The Statistician* 46 (1997), pp. 261–283 (cit. on p. 5-39).
- [2] Kouhei Akazawa, Tsuyoshi Nakamura, and Yuko Palesch. "Power of Logrank Test and Cox Regression Model in Clinical Trials with Heterogeneous Samples". In: *Stat Med* 16 (1997), pp. 583–597 (cit. on p. 13-7).
- [3] Constantin F. Aliferis et al. "Factors Influencing the Statistical Power of Complex Data Analysis Protocols for Molecular Signature Development from Microarray Data". In: *PLoS ONE* 4.3 (2009) (cit. on p. 18-19).
refutation of mic05pre
- [4] D. G. Altman and J. M. Bland. "Absence of Evidence Is Not Evidence of Absence". In: *BMJ* 311 (1995), p. 485 (cit. on p. 20-7).
- [5] Gareth Ambler, Anthony R. Brady, and Patrick Royston. "Simplifying a Prognostic Model: A Simulation Study Based on Clinical Data." In: *Stat Med* 21.24 (Dec. 2002), pp. 3803–3822. issn: 0277-6715. doi: [10.1002/sim.1422](https://doi.org/10.1002/sim.1422). pmid: 12483768. url: <http://dx.doi.org/10.1002/sim.1422> (cit. on p. 20-9).
ordinary backward stepdown worked well when there was a large fraction of truly irrelevant predictors
- [6] Christophe Ambroise and Geoffrey J. McLachlan. "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data". In: *PNASs* 99.10 (May 2002), pp. 6562–6566. issn: 1091-6490. doi: [10.1073/pnas.102102699](https://doi.org/10.1073/pnas.102102699). pmid: 11983868. url: <http://dx.doi.org/10.1073/pnas.102102699> (cit. on p. 20-4).
Relied on an improper accuracy score (proportion classified correct) so had to use the .632 bootstrap unnecessarily
- [7] Per K. Andersen, John P. Klein, and Mei-Jie Zhang. "Testing for Centre Effects in Multi-Centre Survival Studies: A Monte Carlo Comparison of Fixed and Random Effects Tests". In: *Stat Med* 18 (1999), pp. 1489–1500 (cit. on p. 13-33).
- [8] William R. Best et al. "Development of a Crohn's Disease Activity Index". In: *Gastroent* 70 (1976), pp. 439–444 (cit. on p. 18-30).
development of CDAI
- [9] J. Martin Bland and Douglas G. Altman. "Comparisons against Baseline within Randomised Groups Are Often Used and Can Be Highly Misleading". In: *Trials* 12.1 (Dec. 2011), p. 264. doi: [10.1186/1745-6215-12-264](https://doi.org/10.1186/1745-6215-12-264). url: <https://doi.org/10.1186/1745-6215-12-264> (cit. on pp. 13-2, 14-9).
- [10] Robert Bordley. "Statistical Decisionmaking without Math". In: *Chance* 20.3 (2007), pp. 39–44.
- [11] Scott R. Brazer et al. "Using Ordinal Logistic Regression to Estimate the Likelihood of Colorectal Neoplasia". In: *J Clin Epi* 44 (1991), pp. 1263–1270 (cit. on p. 19-13).
- [12] William M. Briggs and Russell Zaretzki. "The Skill Plot: A Graphical Technique for Evaluating Continuous Diagnostic Tests (with Discussion)". In: *Biometrics* 64 (2008), pp. 250–261 (cit. on p. 19-7).
"statistics such as the AUC are not especially relevant to someone who must make a decision about a particular x_c. ... ROC curves lack or obscure several quantities that are necessary for evaluating the operational effectiveness of diagnostic tests. ... ROC curves were first used to check how radio <i>receivers</i> (like radar receivers) operated over a range of frequencies. ... This is not how most ROC curves are used now, particularly in medicine. The receiver of a diagnostic measurement ... wants to make a decision based on some x_c, and is not especially interested in how well he would have done had he used some different cutoff."; in the discussion David Hand states "when integrating to yield the overall AUC measure, it is necessary to decide what weight to give each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data. This is nonsensical. The relative importance of misclassifying a case as a noncase, compared to the reverse, cannot come from the data itself. It must come externally, from considerations of the severity one attaches to the different kinds of misclassifications."; see Lin, Kvam, Lu Stat in Med 28:798-813;2009

- [13] R. M. Califf et al. "The Evolution of Medical and Surgical Therapy for Coronary Artery Disease". In: *JAMA* 261 (1989), pp. 2077–2086 (cit. on p. 13-13).
- [14] R. M. Califf et al. "Prognostic Implications of Ventricular Arrhythmias during 24 Hour Ambulatory Monitoring in Patients Undergoing Cardiac Catheterization for Coronary Artery Disease". In: *Am J Card* 50 (1982), pp. 23–31 (cit. on p. 18-28).
- [15] Mark Chang. *Principles of Scientific Methods*. Chapman and Hall/CRC, Apr. 19, 2016. isbn: 978-0-429-17190-1. doi: [10.1201/b17167](https://doi.org/10.1201/b17167). url: <https://www.taylorfrancis.com/books/9780429171901> (visited on 10/09/2019) (cit. on p. 3-6).
- [16] Qingxia Chen et al. "Too Many Covariates and Too Few Cases? - A Comparative Study". In: *Stat Med* 35.25 (Nov. 2016), pp. 4546–4558. issn: 02776715. doi: [10.1002/sim.7021](https://doi.org/10.1002/sim.7021). url: <http://dx.doi.org/10.1002/sim.7021> (cit. on p. 17-9).
- [17] Leena Choi, Jeffrey D. Blume, and William D. Dupont. "Elucidating the Foundations of Statistical Inference with 2 x 2 Tables". In: *PLoS ONE* 10.4 (Apr. 2015), e0121263+. doi: [10.1371/journal.pone.0121263](https://doi.org/10.1371/journal.pone.0121263). url: <http://dx.doi.org/10.1371/journal.pone.0121263> (cit. on p. 6-6).
- [18] William S. Cleveland. "Graphs in Scientific Publications". In: *Am Statistician* 38 (1984). C/R 85v39 p238-9, pp. 261–269 (cit. on p. 4-10).
- [19] William S. Cleveland. *The Elements of Graphing Data*. Summit, NJ: Hobart Press, 1994 (cit. on p. 4-10).
- [20] T. J. Cole. "Sympercents: Symmetric Percentage Differences on the 100 Log e Scale Simplify the Presentation of Log Transformed Data". In: *Stat Med* 19 (2000), pp. 3109–3125.
- [21] Committee for Proprietary Medicinal Products. "Points to Consider on Adjustment for Baseline Covariates". In: *Stat Med* 23 (2004), pp. 701–709 (cit. on p. 13-34).
- [22] Richard J. Cook and Vern T. Farewell. "Multiplicity Considerations in the Design and Analysis of Clinical Trials". In: *J Roy Stat Soc A* 159 (1996), pp. 93–110 (cit. on p. 5-74).
argues that if results are intended to be interpreted marginally, there may be no need for controlling experimentwise error rate
- [23] Charles S. Davis. *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer, 2002 (cit. on p. 4-11).
- [24] Peter J. Diggle et al. *Analysis of Longitudinal Data*. second. Oxford UK: Oxford University Press, 2002 (cit. on p. 15-5).
- [25] William D. Dupont. *Statistical Modeling for Biomedical Researchers*. second. Cambridge, UK: Cambridge University Press, 2008 (cit. on pp. 15-5, 15-6, 15-14).
- [26] David Edwards. "On Model Pre-Specification in Confirmatory Randomized Studies". In: *Stat Med* 18 (1999), pp. 771–785 (cit. on pp. 13-34, 13-35).
- [27] Bradley Efron and Carl Morris. "Stein's Paradox in Statistics". In: *Sci Am* 236.5 (1977), pp. 119–127 (cit. on p. 14-19).
- [28] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. "Consequences of Dichotomization". In: *Pharm Stat* 8 (2009), pp. 50–61. doi: [10.1002/pst.331](https://doi.org/10.1002/pst.331). url: <http://dx.doi.org/10.1002/pst.331> (cit. on p. 18-15).
optimal cutpoint depends on unknown parameters;should only entertain dichotomization when "estimating a value of the cumulative distribution and when the assumed model is very different from the true model";nice graphics
- [29] Ian Ford, John Norrie, and Susan Ahmadi. "Model Inconsistency, Illustrated by the Cox Proportional Hazards Model". In: *Stat Med* 14 (1995), pp. 735–746 (cit. on p. 13-7).
- [30] M. H. Gail, S. Wieand, and S. Piantadosi. "Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates". In: *Biometrika* 71 (1984), pp. 431–444 (cit. on p. 13-6).
bias if omitted covariables and model is nonlinear
- [31] Mitchell H. Gail. "Adjusting for Covariates That Have the Same Distribution in Exposed and Unexposed Cohorts". In: *Modern Statistical Methods in Chronic Disease Epidemiology*. Ed. by S. H. Moolgavkar and R. L. Prentice. New York: Wiley, 1986, pp. 3–18 (cit. on p. 13-6).
unadjusted test can have larger type I error than nominal

- [32] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 1st ed. Cambridge University Press, Dec. 2006. isbn: 0-521-68689-X. url: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/052168689X> (cit. on pp. 17-3, 17-10).
- [33] A. Giannoni et al. "Do Optimal Prognostic Thresholds in Continuous Physiological Variables Really Exist? Analysis of Origin of Apparent Thresholds, with Systematic Review for Peak Oxygen Consumption, Ejection Fraction and BNP". In: *PLoS ONE* 9.1 (2014). doi: [10.1371/journal.pone.0081699](https://doi.org/10.1371/journal.pone.0081699). url: <http://dx.doi.org/10.1371/journal.pone.0081699> (cit. on pp. 18-11, 18-12).
- [34] David J. Glass. *Experimental Design for Biologists*. 2 edition. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press, Aug. 6, 2014. 294 pp. isbn: 978-1-62182-041-3 (cit. on p. 3-6).
- [35] Tilmann Gneiting and Adrian E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation". In: *J Am Stat Assoc* 102 (2007), pp. 359–378 (cit. on p. 19-6).
wonderful review article except missing references from Scandinavian and German medical decision making literature
- [36] Tim M. Govers et al. "Integrated Prediction and Decision Models Are Valuable in Informing Personalized Decision Making". In: *Journal of Clinical Epidemiology* 0.0 (Aug. 28, 2018). issn: 0895-4356, 1878-5921. doi: [10.1016/j.jclinepi.2018.08.016](https://doi.org/10.1016/j.jclinepi.2018.08.016). pmid: [30170106](https://pubmed.ncbi.nlm.nih.gov/30170106/), [30170106](https://pubmed.ncbi.nlm.nih.gov/30170106/). url: [https://www.jclinepi.com/article/S0895-4356\(18\)30447-5/abstract](https://www.jclinepi.com/article/S0895-4356(18)30447-5/abstract) (visited on 09/01/2018) (cit. on p. 19-8).
- [37] Sander Greenland. "When Should Epidemiologic Regressions Use Random Coefficients?" In: *Biometrics* 56 (2000), pp. 915–921. doi: [10.1111/j.0006-341X.2000.00915.x](https://doi.org/10.1111/j.0006-341X.2000.00915.x). url: <http://dx.doi.org/10.1111/j.0006-341X.2000.00915.x> (cit. on p. 18-4).
use of statistics in epidemiology is largely primitive; stepwise variable selection on confounders leaves important confounders uncontrolled; composition matrix; example with far too many significant predictors with many regression coefficients absurdly inflated when overfit; lack of evidence for dietary effects mediated through constituents; shrinkage instead of variable selection; larger effect on confidence interval width than on point estimates with variable selection; uncertainty about variance of random effects is just uncertainty about prior opinion; estimation of variance is pointless; instead the analysis should be repeated using different values; "if one feels compelled to estimate τ^2 , I would recommend giving it a proper prior concentrated amount contextually reasonable values"; claim about ordinary MLE being unbiased is misleading because it assumes the model is correct and is the only model entertained; shrinkage towards compositional model; "models need to be complex to capture uncertainty about the relations... an honest uncertainty assessment requires parameters for all effects that we know may be present. This advice is implicit in an antiparsimony principle often attributed to L. J. Savage 'All models should be as big as an elephant (see Draper, 1995)'". See also gus06per.
- [38] Sander Greenland et al. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations". In: *Eur J Epi* 31.4 (2016), pp. 337–350. doi: [10.1007/s10654-016-0149-3](https://doi.org/10.1007/s10654-016-0149-3). url: <http://dx.doi.org/10.1007/s10654-016-0149-3> (cit. on p. 5-15).
Best article on misinterpretation of p-values. Pithy summaries
- [39] D. G. Hackam and D. A. Redelmeier. "Translation of Research Evidence from Animals to Humans". In: *JAMA* 296 (2006), pp. 1731–1732 (cit. on p. 21-3).
review of basic science literature that documents systemic methodologic shortcomings. In a personal communication on 20Oct06 the authors reported that they found a few more biostatistical problems that could not make it into the JAMA article (for space constraints); none of the articles contained a sample size calculation; none of the articles identified a primary outcome measure; none of the articles mentioned whether they tested assumptions or did distributional testing (though a few used non-parametric tests); most articles had more than 30 endpoints (but few adjusted for multiplicity, as noted in the article)
- [40] Frank E. Harrell. "Hmisc: A Package of Miscellaneous R Functions". In: (2015). url: <http://biostat.mc.vanderbilt.edu/Hmisc> (cit. on p. 4-5).
- [41] Frank E. Harrell. "Rms: R Functions for Biostatistical/Epidemiologic Modeling, Testing, Estimation, Validation, Graphics, Prediction, and Typesetting by Storing Enhanced Model Design Attributes in the Fit". In: (2016). Implements methods in Regression Modeling Strategies, 2nd edition. url: <http://biostat.mc.vanderbilt.edu/rms> (cit. on pp. 4-28, 9-2).
- [42] Frank E. Harrell et al. "Development of a Clinical Prediction Model for an Ordinal Outcome: The World Health Organization ARI Multicentre Study of Clinical Signs and Etiologic Agents of Pneumonia, Sepsis, and Meningitis in Young Infants". In: *Stat Med* 17 (1998), pp. 909–944. url: [http://onlinelibrary.wiley.com/doi/10.1002/\(SICI\)1097-0258\(19980430\)17:8%3C909::AID-SIM753%3E3.0.CO;2-0/abstract](http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19980430)17:8%3C909::AID-SIM753%3E3.0.CO;2-0/abstract) (cit. on p. 18-8).
- [43] Walter W. Hauck, Sharon Anderson, and Sue M. Marcus. "Should We Adjust for Covariates in Nonlinear Regression Analyses of Randomized Trials?" In: *Controlled Clin Trials* 19 (1998), pp. 249–256. doi: [10.1016/S0197-2456\(97\)00147-5](https://doi.org/10.1016/S0197-2456(97)00147-5). url: [http://dx.doi.org/10.1016/S0197-2456\(97\)00147-5](http://dx.doi.org/10.1016/S0197-2456(97)00147-5) (cit. on pp. 13-9, 13-10).

"For use in a clinician-patient context, there is only a single person, that patient, of interest. The subject-specific measure then best reflects the risks or benefits for that patient. Gail has noted this previously [ENAR Presidential Invited Address, April 1990], arguing that one goal of a clinical trial ought to be to predict the direction and size of a treatment benefit for a patient with specific covariate values. In contrast, population-averaged estimates of treatment effect compare outcomes in groups of patients. The groups being compared are determined by whatever covariates are included in the model. The treatment effect is then a comparison of average outcomes, where the averaging is over all omitted covariates."

- [44] M. A. Hlatky et al. "Criteria for Evaluation of Novel Markers of Cardiovascular Risk: A Scientific Statement from the American Heart Association". In: *Circ* 119.17 (2009). American Heart Association Expert Panel on Subclinical Atherosclerotic Diseases and Emerging Risk Factors and the Stroke Council, pp. 2408–2416 (cit. on p. 19-15). graph with different symbols for diseased and non-diseased

- [45] M. A. Hlatky et al. "Factors Affecting the Sensitivity and Specificity of Exercise Electrocardiography. Multivariable Analysis". In: *Am J Med* 77 (1984), pp. 64–71. url: <http://www.sciencedirect.com/science/article/pii/0002934384904376#> (cit. on p. 19-6).

- [46] Wassily Hoeffding. "A Class of Statistics with Asymptotically Normal Distributions". In: *Ann Math Stat* 19 (1948), pp. 293–325 (cit. on p. 16-2).

Partially reprinted in: Kotz, S., Johnson, N.L. (1992) Breakthroughs in Statistics, Vol I, pp 308-334. Springer-Verlag. ISBN 0-387-94037-5

- [47] Norbert Holländer, Willi Sauerbrei, and Martin Schumacher. "Confidence Intervals for the Effect of a Prognostic Factor after Selection of an 'optimal' Cutpoint". In: *Stat Med* 23 (2004), pp. 1701–1713. doi: [10.1002/sim.1611](https://doi.org/10.1002/sim.1611). url: <http://dx.doi.org/10.1002/sim.1611> (cit. on p. 18-10).

true type I error can be much greater than nominal level;one example where nominal is 0.05 and true is 0.5;minimum P-value method;CART;recursive partitioning;bootstrap method for correcting confidence interval;based on heuristic shrinkage coefficient;"It should be noted, however, that the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the 'optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were included with 12 different cutpoints ... Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology."; dichotomization; categorizing continuous variables; refs alt94dan, sch94out, alt98sub

- [48] Stephen B Hulley et al. *Designing Clinical Research*. Fourth edition. Philadelphia: LWW, July 10, 2013. 378 pp. isbn: 978-1-60831-804-9 (cit. on pp. 3-6, 5-77).

- [49] Cast Investigators. "Preliminary Report: Effect of Encainide and Flecainide on Mortality in a Randomized Trial of Arrhythmia Suppression after Myocardial Infarction". In: *NEJM* 321.6 (1989), pp. 406–412 (cit. on p. 18-29).

- [50] John P. A. Ioannidis and Joseph Lau. "The Impact of High-Risk Patients on the Results of Clinical Trials". In: *J Clin Epi* 50 (1997), pp. 1089–1098. doi: [10.1016/S0895-4356\(97\)00149-2](https://doi.org/10.1016/S0895-4356(97)00149-2). url: [http://dx.doi.org/10.1016/S0895-4356\(97\)00149-2](http://dx.doi.org/10.1016/S0895-4356(97)00149-2) (cit. on p. 13-5).

high risk patients can dominate clinical trials results;high risk patients may be imbalanced even if overall study is balanced;magnesium;differential treatment effect by patient risk;GUSTO;small vs. large trials vs. meta-analysis

- [51] John P. A. Ioannidis. "Expectations, Validity, and Reality in Omics". In: *J Clin Epi* 63 (2010), pp. 945–949 (cit. on p. 21-5).

"Each new field has a rapid exponential growth of its literature over 5–8 years ('new field phase'), followed by an 'established field' phase when growth rates are more modest, and then an 'over-maturity' phase, where the rates of growth are similar to the growth of the scientific literature at large or even smaller. There is a parallel in the spread of an infectious epidemic that emerges rapidly and gets established when a large number of scientists (and articles) are infected with these concepts. Then momentum decreases, although many scientists remain infected and continue to work on this field. New omics infections continuously arise in the scientific community.";"A large number of personal genomic tests are already sold in the market, mostly with direct to consumer advertisement and for 'recreational genomics' purposes (translate: information for the fun of information)."

- [52] Lee Kaiser. "Adjusting for Baseline: Change or Percentage Change?" In: *Stat Med* 8 (1989), pp. 1183–1190. doi: [10.1002/sim.4780081002](https://doi.org/10.1002/sim.4780081002). url: <http://dx.doi.org/10.1002/sim.4780081002>.

- [53] David M. Kent and Rodney Hayward. "Limitations of Applying Summary Results of Clinical Trials to Individual Patients". In: *JAMA* 298 (2007), pp. 1209–1212. doi: [10.1001/jama.298.10.1209](https://doi.org/10.1001/jama.298.10.1209). url: <http://dx.doi.org/10.1001/jama.298.10.1209> (cit. on p. 13-30).

variation in absolute risk reduction in RCTs;failure of subgroup analysis;covariate adjustment;covariate adjustment;nice summary of individual patient absolute benefit vs. patient risk

- [54] William A. Knaus et al. "The Clinical Evaluation of New Drugs for Sepsis: A Prospective Study Design Based on Survival Analysis". In: *JAMA* 270 (1993), pp. 1233–1241. doi: [10.1001/jama.270.10.1233](https://doi.org/10.1001/jama.270.10.1233). url: <http://dx.doi.org/10.1001/jama.270.10.1233> (cit. on p. 13-34).
- [55] R. Koenker and G. Bassett. "Regression Quantiles". In: *Econometrica* 46 (1978), pp. 33–50 (cit. on p. 19-16).
- [56] Samuel Kotz and Norman L. Johnson, eds. *Encyclopedia of Statistical Sciences*. Vol. 9. New York: Wiley, 1988 (cit. on p. 3-20).
- [57] R. A. Kronmal. "Spurious Correlation and the Fallacy of the Ratio Standard Revisited". In: *J Roy Stat Soc A* 156 (1993), pp. 379–392.
- spurious correlation in using ratio variables even if all component variables of ratios are uncorrelated; division of only the dependent variable by an independent variable can result in regression coefficient estimates for the other independent variables that result in inappropriate conclusions; use of a ratio as an independent variable can result in inadequate adjustment for component variables of the ratio; ratio variables should only be used in a full model containing all the component variables; results of regression analyses incorporating ratios are not readily comparable across studies
- [58] Jan S. Krouwer. "Why Bland-Altman Plots Should Use X, Not (Y+X)/2 When X Is a Reference Method". In: *Stat Med* 27.5 (Feb. 2008), pp. 778–780. issn: 0277-6715. doi: [10.1002/sim.3086](https://doi.org/10.1002/sim.3086). pmid: [17907247](https://pubmed.ncbi.nlm.nih.gov/17907247/). url: <http://dx.doi.org/10.1002/sim.3086> (cit. on p. 8-11).
- [59] Jeffery T. Leek and Roger D. Peng. "What Is the Question?" In: *Science* 347.6228 (Mar. 20, 2015), pp. 1314–1315. issn: 0036-8075, 1095-9203. doi: [10.1126/science.aaa6146](https://doi.org/10.1126/science.aaa6146). pmid: [25721505](https://pubmed.ncbi.nlm.nih.gov/25721505/). url: <https://science.sciencemag.org/content/347/6228/1314> (visited on 09/27/2019) (cit. on p. 3-8).
- [60] Russell V. Lenth. "Some Practical Guidelines for Effective Sample Size Determination". In: *Am Statistician* 55 (2001), pp. 187–193. doi: [10.1198/000313001317098149](https://doi.org/10.1198/000313001317098149). url: <http://dx.doi.org/10.1198/000313001317098149> (cit. on p. 5-72).
- problems with Cohen's method
- [61] R. J. MacKay and R. W. Oldford. "Scientific Method, Statistical Method and the Speed of Light". In: *Statist. Sci.* 15.3 (Aug. 1, 2000), pp. 254–278. issn: 0883-4237, 2168-8745. doi: [10.1214/ss/1009212817](https://doi.org/10.1214/ss/1009212817). url: <https://projecteuclid.org/euclid.ss/1009212817> (visited on 09/16/2019) (cit. on p. 3-5).
- [62] J. S. Maritz. "Models and the Use of Signed Rank Tests". In: *Stat Med* 4 (1985), pp. 145–153. doi: [10.1002/sim.4780040205](https://doi.org/10.1002/sim.4780040205). url: <http://dx.doi.org/10.1002/sim.4780040205>.
- [63] J. N. S. Matthews and N. H. Badi. "Inconsistent Treatment Estimates from Mis-Specified Logistic Regression Analyses of Randomized Trials". In: *Stat Med* 34.19 (Aug. 2015), pp. 2681–2694. issn: 02776715. doi: [10.1002/sim.6508](https://doi.org/10.1002/sim.6508). url: <http://dx.doi.org/10.1002/sim.6508> (cit. on p. 13-3).
- [64] J. N. S. Matthews et al. "Analysis of Serial Measurements in Medical Research". In: *BMJ* 300 (1990). Letter to editor by S. Senn in same issue, pp. 230–235. doi: [10.1136/bmj.300.6719.230](https://doi.org/10.1136/bmj.300.6719.230). url: <http://dx.doi.org/10.1136/bmj.300.6719.230> (cit. on p. 15-5).
- [65] Stefan Michiels, Serge Koscielny, and Catherine Hill. "Prediction of Cancer Outcome with Microarrays: A Multiple Random Validation Strategy". In: *Lancet* 365 (2005), pp. 488–492 (cit. on p. 18-19).
- comment on p. 454; validation; microarray; bioinformatics; machine learning; nearest centroid; severe problems with data splitting; high variability of list of genes; problems with published studies; nice results for effect of training sample size on misclassification error; nice use of confidence intervals on accuracy estimates; unstable molecular signatures; high instability due to dependence on selection of training sample
- [66] Karel G. M. Moons and Frank E. Harrell. "Sensitivity and Specificity Should Be De-Emphasized in Diagnostic Accuracy Studies". In: *Acad Rad* 10 (2003). Editorial, pp. 670–672 (cit. on p. 19-6).
- [67] Karel G. M. Moons et al. "Limitations of Sensitivity, Specificity, Likelihood Ratio, and Bayes' Theorem in Assessing Diagnostic Probabilities: A Clinical Example". In: *Epi* 8.1 (1997), pp. 12–17 (cit. on p. 19-6).
- non-constancy of sensitivity, specificity, likelihood ratio in a real example
- [68] Thomas J. Moore. *Deadly Medicine: Why Tens of Thousands of Patients Died in America's Worst Drug Disaster*. review in *Stat in Med* 16:2507-2510, 1997. New York: Simon & Shuster, 1995 (cit. on pp. 18-25, 18-26, 18-29).
- [69] Multicenter Postinfarction Research Group. "Risk Stratification and Survival after Myocardial Infarction". In: *NEJM* 309 (1983), pp. 331–336 (cit. on pp. 18-26, 18-27, 18-28).
- terrible example of dichotomizing continuous variables; figure ins Papers/modelingPredictors

- [70] Marcus R. Munafò et al. "A Manifesto for Reproducible Science". In: *Nat Hum Behav* 1.1 (Jan. 2017), pp. 0021+. issn: 2397-3374. doi: [10.1038/s41562-016-0021](https://doi.org/10.1038/s41562-016-0021). url: <http://dx.doi.org/10.1038/s41562-016-0021> (cit. on p. 21-1).
- [71] Paul Murrell. "InfoVis and Statistical Graphics: Comment". In: *J Comp Graph Stat* 22.1 (2013), pp. 33–37. doi: [10.1080/10618600.2012.751875](https://doi.org/10.1080/10618600.2012.751875). eprint: <http://www.tandfonline.com/doi/pdf/10.1080/10618600.2012.751875>. url: <http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.751875> (cit. on p. 4-10).
- Excellent brief how-to list; incorporated into graphscourse
- [72] O. Naggara et al. "Analysis by Categorizing or Dichotomizing Continuous Variables Is Inadvisable: An Example from the Natural History of Unruptured Aneurysms". In: *Am J Neuroradiol* 32.3 (2011), pp. 437–440. doi: [10.3174/ajnr.A2425](https://doi.org/10.3174/ajnr.A2425). url: <http://www.ajnr.org/content/32/3/437.abstract> (cit. on p. 18-12).
- [73] John M. Neuhaus. "Estimation Efficiency with Omitted Covariates in Generalized Linear Models". In: *J Am Stat Assoc* 93 (1998), pp. 1124–1129 (cit. on p. 13-10).
- "to improve the efficiency of estimated covariate effects of interest, analysts of randomized clinical trial data should adjust for covariates that are strongly associated with the outcome, and ... analysts of observational data should not adjust for covariates that do not confound the association of interest"
- [74] Thomas B. Newman and Michael A. Kohn. *Evidence-Based Diagnosis*. Cambridge University Press, 2009. 312 pp. isbn: 978-0-521-71402-0 (cit. on p. 19-5).
- [75] Regina Nuzzo. "How Scientists Fool Themselves — and How They Can Stop". In: *Nature* 526.7572 (Oct. 2015), pp. 182–185 (cit. on pp. 21-1, 21-8, 21-9).
- [76] Peter C. O'Brien. "Comparing Two Samples: Extensions of the t, Rank-Sum, and Log-Rank Test". In: *J Am Stat Assoc* 83 (1988), pp. 52–61. doi: [10.1080/01621459.1988.10478564](https://doi.org/10.1080/01621459.1988.10478564). url: http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478564#U_sn0XVdU3E (cit. on p. 15-10).
- see Hauck WW, Hyslop T, Anderson S (2000) Stat in Med 19:887-899
- [77] E. M. Ohman et al. "Cardiac Troponin T Levels for Risk Stratification in Acute Myocardial Ischemia". In: *NEJM* 335 (1996), pp. 1333–1341 (cit. on p. 18-4).
- [78] Guillaume Paré, Shamir R. Mehta, Salim Yusuf, et al. "Effects of CYP2C19 Genotype on Outcomes of Clopidogrel Treatment". In: *NEJM* online (2010). August 29, 2010 (cit. on p. 10-51).
- [79] Debasish Paul et al. "'Preconditioning' for Feature Selection and Regression in High-Dimensional Problems". In: *Ann Stat* 36.4 (2008), pp. 1595–1619. doi: [10.1214/009053607000000578](https://doi.org/10.1214/009053607000000578). url: <http://dx.doi.org/10.1214/009053607000000578> (cit. on p. 20-9).
- develop consistent Y using a latent variable structure, using for example supervised principal components. Then run stepwise regression or lasso predicting Y (lasso worked better). Can run into problems when a predictor has importance in an adjusted sense but has no marginal correlation with Y;model approximation;model simplification
- [80] Michael J. Pencina et al. "Evaluating the Added Predictive Ability of a New Marker: From Area under the ROC Curve to Reclassification and Beyond". In: *Stat Med* 27 (2008), pp. 157–172 (cit. on p. 19-14).
- small differences in ROC area can still be very meaningful;example of insignificant test for difference in ROC areas with very significant results from new method;Yates' discrimination slope;reclassification table;limiting version of this based on whether and amount by which probabilities rise for events and lower for non-events when compare new model to old;comparing two models;see letter to the editor by Van Calster and Van Huffel, Stat in Med 29:318-319, 2010 and by Cook and Paynter, Stat in Med 31:93-97, 2012
- [81] John R. Platt. "Strong Inference". In: *Science* 146.3642 (1964), pp. 347–353 (cit. on p. 21-8).
- [82] David B. Pryor et al. "Estimating the Likelihood of Significant Coronary Artery Disease". In: *Am J Med* 75 (1983), pp. 771–780 (cit. on p. 19-11).
- [83] Gillian M. Raab, Simon Day, and Jill Sales. "How to Select Covariates to Include in the Analysis of a Clinical Trial". In: *Controlled Clin Trials* 21 (2004), pp. 330–342 (cit. on p. 13-10).
- how correlated with outcome must a variable before adding it helps more than hurts, as a function of sample size;planning;design;variable selection
- [84] L. D. Robinson and N. P. Jewell. "Some Surprising Results about Covariate Adjustment in Logistic Regression Models". In: *Int Stat Rev* 59 (1991), pp. 227–240 (cit. on p. 13-3).

- [85] Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. "Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea". In: *Stat Med* 25 (2006), pp. 127–141. doi: [10.1002/sim.2331](https://doi.org/10.1002/sim.2331). url: <http://dx.doi.org/10.1002/sim.2331> (cit. on pp. 18-10, 18-12).
- destruction of statistical inference when cutpoints are chosen using the response variable; varying effect estimates when change cut-points; difficult to interpret effects when dichotomize; nice plot showing effect of categorization; PBC data
- [86] Donald B. Rubin. "The Design versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Studies". In: *Stat Med* 26 (2007), pp. 20–36 (cit. on p. 21-9).
- [87] Ruxton, Graeme D. and Colegrave, Nick. *Experimental Design for the Life Sciences*. Fourth Edition. Oxford, New York: Oxford University Press, Oct. 15, 2017. 208 pp. isbn: 978-0-19-871735-5 (cit. on p. 3-6).
- [88] Daniel J. Sargent and James S. Hodges. "A Hierarchical Model Method for Subgroup Analysis of Time-to-Event Data in the Cox Regression Setting". Presented at the Joint Statistical Meetings, Chicago; see and99tes. 1996 (cit. on p. 13-12).
- [89] David A. Schoenfeld. "Sample Size Formulae for the Proportional Hazards Regression Model". In: *Biometrics* 39 (1983), pp. 499–503 (cit. on p. 13-8).
- [90] Greg Schwemer. "General Linear Models for Multicenter Clinical Trials". In: *Controlled Clin Trials* 21 (2000), pp. 21–29 (cit. on p. 13-33).
- [91] Stephen Senn. "Controversies Concerning Randomization and Additivity in Clinical Trials". In: *Stat Med* 23 (2004), pp. 3729–3753. doi: [10.1002/sim.2074](https://doi.org/10.1002/sim.2074). url: <http://dx.doi.org/10.1002/sim.2074> (cit. on pp. 13-6, 13-36, 13-37).
- p. 3735: "in the pharmaceutical industry, in analyzing the data, if a linear model is employed, it is usual to fit centre as a factor but unusual to fit block."; p. 3739: a large trial "is not less vulnerable to chance covariate imbalance"; p. 3741: "There is no place, in my view, for classical minimization" (vs. the method of Atkinson); "If an investigator uses such [allocation based on covariates] schemes, she or he is honour bound, in my opinion, as a very minimum, to adjust for the factors used to balance, since the fact that they are being used to balance is an implicit declaration that they will have prognostic value."; "The point of view is sometimes defended that analyses that ignore covariates are superior because they are simpler. I do not accept this. A value of $\pi=3$ is a simple one and accurate to one significant figure ... However very few would seriously maintain that it should generally be adopted by engineers."; p. 3742: "as Fisher pointed out ... if we balance by a predictive covariate but do not fit the covariate in the model, not only do we not exploit the covariate, we actually increase the expected declared standard error."; p. 3744: "I would like to see standard errors for group means abolished."; p. 3744: "A common habit, however, in analyzing trials with three or more arms is to pool the variances from all arms when calculating the standard error of a given contrast. In my view this is a curious practice ... it relies on an assumption of additivity of $\sum a_i \pi_i$ treatments when comparing only $\sum a_i \pi_i$... a classical t-test is robust to heteroscedasticity provided that sample sizes are equal in the groups being compared and that the variance is internal to those two groups but is not robust where an external estimate is being used."; p. 3745: "By adjusting main effects for interactions a type III analysis is similarly illogical to Neyman's hypothesis test."; "Guyatt et al. ... found a 'method for estimating the proportion of patients who benefit from a treatment ... In fact they had done no such thing.'"; p. 3746: "When I checked the Web of Science on 29 June 2003, the paper by Horwitz et al. had been cited 28 times and that by Guyatt et al. had been cited 79 times. The letters pointing out the fallacies had been cited only 8 and 5 times respectively."; "if we pool heterogeneous strata, the odds ratio of the treatment effect will be different from that in every stratum, even if from stratum to stratum it does not vary."; p. 3747: "Part of the problem with Poisson, proportional hazard and logistic regression approaches is that they use a single parameter, the linear predictor, with no equivalent of the variance parameter in the Normal case. This means that lack of fit impacts on the estimate of the predictor. ... what is the value of randomization if, in all except the Normal case, we cannot guarantee to have unbiased estimates. My view ... was that the form of analysis envisaged (that is to say, which factors and covariates should be fitted) justified the allocation and $\sum a_i \pi_i$ not vice versa."; "use the additive measure at the point of analysis and transform to the relevant scale at the point of implementation. This transformation at the point of medical decision-making will require auxiliary information on the level of background risk of the patient."; p. 3748: "The decision to fit prognostic factors has a far more dramatic effect on the precision of our inferences than the choice of an allocation based on covariates or randomization approach and one of my chief objections to the allocation based on covariates approach is that trialists have tended to use the fact that they have balanced as an excuse for not fitting. This is a grave mistake."
- [92] Stephen Senn. *Statistical Issues in Drug Development*. Second. Chichester, England: Wiley, 2008 (cit. on pp. 14-9, 18-15).
- [93] Stephen J. Senn. "Dichotomania: An Obsessive Compulsive Disorder That Is Badly Affecting the Quality of Analysis of Pharmaceutical Trials". In: *Proceedings of the International Statistical Institute, 55th Session*. Sydney, 2005. url: <http://hbiostat.org/papers/Senn/dichotomania.pdf> (cit. on pp. 18-10, 18-14).
- [94] Stephen Senn, Vladimir V. Anisimov, and Valerii V. Fedorov. "Comparisons of Minimization and Atkinson's Algorithm". In: *Stat Med* 29 (2010), pp. 721–730 (cit. on p. 13-36).
- "fitting covariates may make a more valuable and instructive contribution to inferences about treatment effects than only balancing them"
- [95] Stephen Senn, Lynda Stevens, and Nish Chaturvedi. "Repeated Measures in Clinical Trials: Simple Strategies for Analysis Using Summary Measures". In: *Stat Med* 19 (2000), pp. 861–877. doi: [10.1002/\(SICI\)1097-0258\(20000330\)19:6\%3C861::AID-SIM407\%3E3.0.CO;2-F](https://doi.org/10.1002/(SICI)1097-0258(20000330)19:6\%3C861::AID-SIM407\%3E3.0.CO;2-F). url: [http://dx.doi.org/10.1002/\(SICI\)1097-0258\(20000330\)19:6\%3C861::AID-SIM407\%3E3.0.CO;2-F](http://dx.doi.org/10.1002/(SICI)1097-0258(20000330)19:6\%3C861::AID-SIM407\%3E3.0.CO;2-F) (cit. on p. 15-5).

- [96] Xiaotong Shen, Hsin-Cheng Huang, and Jimmy Ye. "Inference after Model Selection". In: *J Am Stat Assoc* 99 (2004), pp. 751–762 (cit. on p. 20-9).
 uses optimal approximation for estimating mean and variance of complex statistics adjusting for model selection
- [97] Arminda L. Sigueira and Jeremy M. G. Taylor. "Treatment Effects in a Logistic Model Involving the Box-Cox Transformation". In: *J Am Stat Assoc* 94 (1999), pp. 240–246 (cit. on p. 13-35).
 Box-Cox transformation of a covariate; validity of inference for treatment effect when treat exponent for covariate as fixed
- [98] Alan Spanos, Frank E. Harrell, and David T. Durack. "Differential Diagnosis of Acute Meningitis: An Analysis of the Predictive Value of Initial Observations". In: *JAMA* 262 (1989), pp. 2700–2707. doi: [10.1001/jama.262.19.2700](https://doi.org/10.1001/jama.262.19.2700). url: <http://dx.doi.org/10.1001/jama.262.19.2700> (cit. on p. 19-12).
- [99] Ewout W. Steyerberg. "Validation in Prediction Research: The Waste by Data-Splitting". In: *Journal of Clinical Epidemiology* 0.0 (July 28, 2018). issn: 0895-4356, 1878-5921. doi: [10.1016/j.jclinepi.2018.07.010](https://doi.org/10.1016/j.jclinepi.2018.07.010). url: [https://www.jclinepi.com/article/S0895-4356\(18\)30485-2/abstract](https://www.jclinepi.com/article/S0895-4356(18)30485-2/abstract) (visited on 07/30/2018) (cit. on p. 10-59).
- [100] Ewout W. Steyerberg, Patrick M. M. Bossuyt, and Kerry L. Lee. "Clinical Trials in Acute Myocardial Infarction: Should We Adjust for Baseline Characteristics?" In: *Am Heart J* 139 (2000). Editorial, pp. 761-763, pp. 745–751. doi: [10.1016/S0002-8703\(00\)90001-2](https://doi.org/10.1016/S0002-8703(00)90001-2). url: [http://dx.doi.org/10.1016/S0002-8703\(00\)90001-2](http://dx.doi.org/10.1016/S0002-8703(00)90001-2) (cit. on p. 13-4).
- [101] Victoria Stodden, Friedrich Leisch, and Roger D. Peng. *Implementing Reproducible Research*. Ed. by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Boca Raton, FL: CRC Press/Taylor and Francis, 2014. isbn: 978-1-4665-6159-5. url: <http://www.worldcat.org/isbn/9781466561595> (cit. on p. 21-16).
- [102] Jyothi Subramanian and Richard Simon. "Gene Expression-Based Prognostic Signatures in Lung Cancer: Ready for Clinical Use?" In: *J Nat Cancer Inst* 102 (2010), pp. 464–474 (cit. on p. 21-4).
 none demonstrated to have clinical utility; bioinformatics; quality scoring of papers
- [103] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2016. isbn: 3-900051-07-0. url: <http://www.R-project.org> (cit. on p. 1-1).
- [104] L. Törnqvist, P. Vartia, and Y. O. Vartia. "How Should Relative Changes Be Measured?" In: *Am Statistician* 39 (1985), pp. 43–46. doi: [10.1080/00031305.1985.10479385](https://doi.org/10.1080/00031305.1985.10479385). url: <http://dx.doi.org/10.1080/00031305.1985.10479385>.
- [105] John W. Tukey. "Tightening the Clinical Trial". In: *Controlled Clin Trials* 14 (1993), pp. 266–285. doi: [10.1016/0197-2456\(93\)90225-3](https://doi.org/10.1016/0197-2456(93)90225-3). url: [http://dx.doi.org/10.1016/0197-2456\(93\)90225-3](http://dx.doi.org/10.1016/0197-2456(93)90225-3) (cit. on p. 13-10).
 showed that asking clinicians to make up regression coefficients out of thin air is better than not adjusting for covariates
- [106] Tjeerd van derPloeg, Peter C. Austin, and Ewout W. Steyerberg. "Modern Modelling Techniques Are Data Hungry: A Simulation Study for Predicting Dichotomous Endpoints." In: *BMC medical research methodology* 14.1 (Dec 2014), pp. 137+. issn: 1471-2288. doi: [10.1186/1471-2288-14-137](https://doi.org/10.1186/1471-2288-14-137). pmid: 25532820. url: <http://dx.doi.org/10.1186/1471-2288-14-137> (cit. on p. 20-8).
 Would be better to use proper accuracy scores in the assessment. Too much emphasis on optimism as opposed to final discrimination measure. But much good practical information. Recursive partitioning fared poorly.
- [107] David van Klaveren et al. "Estimates of Absolute Treatment Benefit for Individual Patients Required Careful Modeling of Statistical Interactions". In: *J Clin Epi* 68.11 (Nov. 2015), pp. 1366–1374. issn: 08954356. doi: [10.1016/j.jclinepi.2015.02.012](https://doi.org/10.1016/j.jclinepi.2015.02.012). url: <http://dx.doi.org/10.1016/j.jclinepi.2015.02.012> (cit. on p. 13-24).
- [108] Andrew J. Vickers. "Decision Analysis for the Evaluation of Diagnostic Tests, Prediction Models, and Molecular Markers". In: *Am Statistician* 62.4 (2008), pp. 314–320.
 limitations of accuracy metrics; incorporating clinical consequences; nice example of calculation of expected outcome; drawbacks of conventional decision analysis, especially because of the difficulty of eliciting the expected harm of a missed diagnosis; use of a threshold on the probability of disease for taking some action; decision curve; has other good references to decision analysis
- [109] Andrew J. Vickers, Ethan Basch, and Michael W. Kattan. "Against Diagnosis". In: *Ann Int Med* 149 (2008), pp. 200–203 (cit. on p. 19-5).

"The act of diagnosis requires that patients be placed in a binary category of either having or not having a certain disease. Accordingly, the diseases of particular concern for industrialized countries—such as type 2 diabetes, obesity, or depression—require that a somewhat arbitrary cut-point be chosen on a continuous scale of measurement (for example, a fasting glucose level $>6.9 \text{ mmol/L}$ [$>125 \text{ mg/dL}$] for type 2 diabetes). These cut-points do not adequately reflect disease biology, may inappropriately treat patients on either side of the cut-point as 2 homogenous risk groups, fail to incorporate other risk factors, and are invariable to patient preference."

- [110] Howard Wainer. "Finding What Is Not There through the Unfortunate Binning of Results: The Mendel Effect". In: *Chance* 19.1 (2006), pp. 49–56 (cit. on p. 18-10).
- can find bins that yield either positive or negative association;especially pertinent when effects are small;"With four parameters, I can fit an elephant; with five, I can make it wiggle its trunk." - John von Neumann
- [111] Ian R. White and Simon G. Thompson. "Adjusting for Partially Missing Baseline Measurements in Randomized Trials". In: *Stat Med* 24 (2005), pp. 993–1007 (cit. on p. 13-35).
- [112] John Whitehead. "Sample Size Calculations for Ordered Categorical Data". In: *Stat Med* 12 (1993). See letter to editor SM 15:1065-6 for binary case;see errata in SM 13:871 1994;see kol95com, jul96sam, pp. 2257–2271 (cit. on p. 7-31).
- [113] Rand Wilcox et al. "Avoid Lost Discoveries, Because of Violations of Standard Assumptions, by Using Modern Robust Statistical Methods". In: *Journal of Clinical Epidemiology* 66.3 (Mar. 1, 2013), pp. 319–329. issn: 0895-4356, 1878-5921. doi: [10.1016/j.jclinepi.2012.09.003](https://doi.org/10.1016/j.jclinepi.2012.09.003). pmid: [23195918](https://pubmed.ncbi.nlm.nih.gov/23195918/). url: [https://www.jclinepi.com/article/S0895-4356\(12\)00275-2/abstract](https://www.jclinepi.com/article/S0895-4356(12)00275-2/abstract) (visited on 10/24/2019) (cit. on p. 5-5).
- [114] Yihui Xie. *Dynamic Documents with R and Knitr, Second Edition*. second. Chapman and Hall, 2015. isbn: 978-1-4987-1696-3 (cit. on p. 21-12).
- [115] Takuhiro Yamaguchi and Yasuo Ohashi. "Investigating Centre Effects in a Multi-Centre Clinical Trial of Superficial Bladder Cancer". In: *Stat Med* 18 (1999), pp. 1961–1971 (cit. on pp. 13-12, 13-33).