

BIOS 662 Fall 2018

Study Designs

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://www.unc.edu/sakai/>

Types of Studies

- Definition 2.1. An *observational study* collects data from an existing situation. The data collection does not intentionally interfere with the running of the system.
→ beware *Hawthorne effect*.
- Definition 2.2. An *experiment* is a study in which an investigator deliberately sets one or more factors to a specified level.
→ leads to stronger scientific/causal inference

Types of Biomedical Studies

- Observational studies:

- Cross-sectional
- Longitudinal
- Cohort
- Case Control

- Experimental Studies:

- Laboratory
- Clinical Trials

Definition 2.5: An *experimental unit* or *study unit* is the smallest unit on which an experiment or study is performed.

→ Distinct from the unit of observation.

Cross-Sectional Studies

- Definition 2.16: A *cross-sectional study* collects data on study units at a (single) fixed time.
- Purposes:

1. To describe a population at a point in time; measure prevalence

Examples:

- U.S. Census
- National Health and Nutrition Examination Surveys (NHANES)

2. To examine associations

Examples:

- menopause status and blood cholesterol level
- hyperactivity and blood lead levels
- diet and blood pressure

Longitudinal Studies

- Definition 2.16: A *longitudinal study* collects information on study units on at least two occasions.
- Purposes:
 1. To measure change

Examples:

- change in height for each year of age
- change in viral load in HIV+ individuals

2. To develop predictions

Example:

- given blood pressure at age 10, what would we expect blood pressure to be at age 15?

3. To examine the association between the changes in 2 or more variables (explore temporal changes)

Example:

- is change in viral load associated with change in CD4+ T-cell count?

Cohort Studies

- Definition 2.10: A *cohort study* or *prospective study* is one in which a cohort of people is identified and followed to observe specified endpoints (e.g., occurrence of disease).
- Purposes:
 1. Estimate incidence of disease
 2. Relate baseline measures to the occurrence of disease

Examples:

- Framingham Heart Study
- Atherosclerosis Risk in Communities (ARIC) Study

Question:

- Are cohort and longitudinal studies mutually exclusive?
- If not, how are they related?

Case-Control Studies

- Definition 2.12: A *retrospective study* is one in which people having a particular outcome or endpoint are identified and studied.
- Definition 2.13: A *case-control study* selects all cases, usually of a (rare) disease, that meet fixed criteria.
A disease-free group, called *controls*, that serve as a comparison for the cases is also selected.
The cases and controls are compared with respect to various characteristics (exposures, risk factors)

Case-Control Studies cont.

- Definition 2.14: In a *matched case-control study*, controls are selected to match characteristics of individual cases.
The cases and controls(s) are associated with each other.
There may be more than one control for each case.
In an *individually-matched case-control study*, each control is matched with a specific case.
- Definition 2.15: In a *frequency-matched case-control study*, controls are selected to match characteristics of the entire case sample. A control is not matched with a specific case.

Case-Control Studies cont.

- Example: Investigators interested in the association between thromboembolic disease and oral contraceptive use
 - Cases: women aged 16-40 who had been discharged from one of 19 hospitals for deep vein thrombosis
 - Controls: women suffering acute medical conditions (other than thromboembosis) or elective surgery
 - Individual matching, with two controls per case; matched on age, date of hospital admission, parity
 - All participants asked about oral contraceptive history (50% cases, 14% controls)

Experimental Studies

- Definition: Interventions are applied by investigator
- Purpose: to compare outcomes between two or more interventions
- Example: Compared to placebo, does a candidate vaccine result in a lower incidence of HIV in high risk individuals?
- Usually interventions are assigned using *randomization*: a random but known process by which participants are assigned to different treatments or interventions
For instance, each participant may be equally likely to be assigned to a medication or matching placebo

Randomization

- Attributed to R. A. Fisher
- “One of the great intellectual advances of the twentieth century.”

Lloyd D. Fisher

- Advantages:
 - Removes potential bias in allocating participants to different intervention groups
 - Tends to produce comparable groups on average
 - Provides a basis for statistical tests
- Disadvantages:
 - Ethical?
- RCT = randomized controlled trial

Clinical Trials

- Treatment trials – pharmaceutical, surgical, etc.
- Prevention trials
- Screening trials
- Diagnostic trials

Pharmaceutical trials

- Pre-clinical
- Phase I
- Phase II
- Phase III
- Phase IV

Parallel Group and Factorial Experiments

Crossover Experiment

- Definition 2.6: In a *crossover experiment* the sample experimental unit receives more than one treatment during non-overlapping time periods.
- Advantage: each experimental unit serves as its own control, eliminating subject-to-subject variability
- Disadvantage: possible carryover effects of treatment, calendar time effects
- Usually randomize the order of the treatments

Blinding / Masking

- Definition 2.19: A study is *single blind* if subjects being treated are unaware of which treatment (including any control) they are receiving
- A study is *double blind* if both the participants and the researchers are unaware of which treatment the subjects are receiving
- *Triple blind*: double blind plus those analyzing or reviewing the data (statistician/monitoring committee) are unaware of treatment assignments
- Sometimes impossible/infeasible: nutrition, circumcision
- Unblinded studies (aka “open label”):
 - Disadvantage: potential for bias (systematic error)
 - Advantages: reflects clinical practice, simpler

Endpoints

- Definition 2.9: An *endpoint* is a clearly defined outcome or event associated with an experimental or study unit
- Important considerations in choosing an endpoint:
 - Relevance
 - Reliability
 - Rate
- Hierarchy: primary, secondary, tertiary, etc.

Steps in Performing a Study

- Identify a question or problem area of interest
- Design a study to answer the question
 - Decide on the type of study
 - Identify the data to be collected
 - Determine appropriate analytical models
 - Determine the sample size required
- Conduct the study and collect the data
- Analyze the data and draw conclusions and inferences
- Use the results

Inferences from a Study

- What was the design?
- Guard against bias:
 - Comparability
 - Representative of target population
- Source of, control for, and quantification of uncertainty/variation

Some Considerations Related to Ethics

- Stakeholders in a study
 - Subjects / patients / participants
 - Population at risk
 - Funding agency
 - Scientific advancement
 - Study investigators
- Institutional Review Boards and external monitoring boards
- Principle of informed consent
- It is unethical to enroll participants in a study that is not designed appropriately to address the question of interest

BIOS 668 and 752

668 DESIGN OF PUBLIC HEALTH STUDIES (3). Prerequisites, BIOS 545, 550, or equivalents. Statistical concepts in basic public health study designs: cross-sectional, case-control, prospective, and experimental (including clinical trials). Validity, measurement of response, sample size determination, matching and random allocation methods. Spring.

752 DESIGN AND ANALYSIS OF CLINICAL TRIALS (3)

Prerequisites, BIOS 660, and 661 or permission of the instructor.

Description: This course will introduce the methods used in clinical trials. Topics include dose-finding trials, allocation to treatments in randomized trials, sample size calculation, interim monitoring, and non-inferiority trials. Fall. Ivanova.

BIOS 662 Fall 2018

Introduction to R

Based on a set of notes by

Wei Sun, Ph.D.

Department of Biostatistics

University of North Carolina at Chapel Hill

What is R?

A language and software environment for statistical computing and graphics.

- R is free!
- It is open-source and involves many developers.
- The R system is developing rapidly.
- Straightforward simple calculations and analysis.
- Allows low level control for some tasks.
- Extensive graphical abilities.
- Sometimes R is slow...

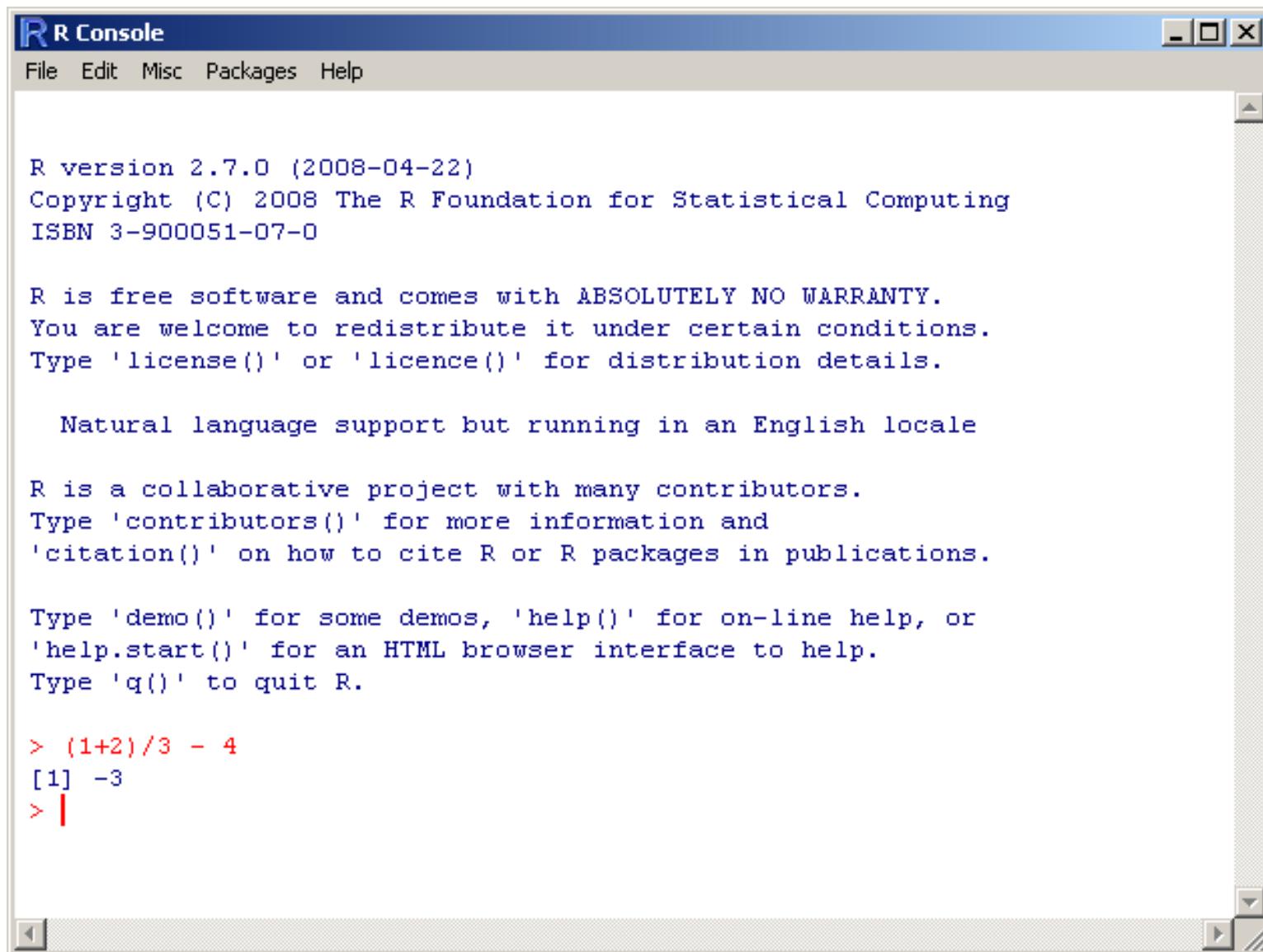


Figure 1: R graphical user interface (Windows)

The image shows two windows side-by-side. On the left is the R Console window, which displays a series of R commands and their results. On the right is the Tinn-R editor window, which shows the same R code in a script file named R_calculator.R. The code in the editor is identical to what is shown in the R Console. The Tinn-R window has a toolbar at the top with various icons for file operations, and the code area has syntax highlighting. A yellow highlight bar is present under the line of code '18 # square root' in the editor.

```
> # How many seconds per year?  
> 60*60*24*365  
[1] 31536000  
>  
> # remainder  
> 82 %% 10  
[1] 2  
>  
> # take log  
> log(100)  
[1] 4.60517  
>  
> # log base 10  
> log10(100)  
[1] 2  
>  
> # exponential  
> exp(4.60517)  
[1] 99.99998  
>  
> # power  
> 2^10  
[1] 1024  
>  
> # square root  
> sqrt(81)  
[1] 9  
>  
> |
```

```
1 # How many seconds per year?  
2 60*60*24*365  
3  
4 # remainder  
5 82 %% 10  
6  
7 # take log  
8 log(100)  
9  
10 # log base 10  
11 log10(100)  
12  
13 # exponential  
14 exp(4.60517)  
15  
16 # power  
17 2^10  
18  
19 # square root  
20 sqrt(81)  
21
```

Figure 2: Use an appropriate editor, e.g., Tinn-R

Using R as a calculator

```
> # How many seconds in a year?  
> 60*60*24*365  
[1] 31536000  
>  
> # remainder  
> 82 %% 10  
[1] 2  
>  
> # natural log and log to base 10  
> log(100)  
[1] 4.60517  
>  
> log10(100)  
[1] 2  
>  
> # exponential  
> exp(4.60517)  
[1] 99.99998
```

```
>  
> # power  
> 2^10  
[1] 1024  
>  
> # square root  
> sqrt(81)  
[1] 9
```

Scalar Variables

```
> # Define a variable
> x = 123.45
> x
[1] 123.45
>
> # R language is case sensitive
> X
Error: object "X" not found
>
> # another way to define a variable
> z <- 66.55
> z + x
[1] 190
>
> # be careful
> w <- 8.9
Error: object "w" not found
```

Vectors

```
> # Define a vector  
> v = c(1.2, 2.3, 3.4)  
> v  
[1] 1.2 2.3 3.4  
> v*2  
[1] 2.4 4.6 6.8  
>  
> # summation  
> sum(v)  
[1] 6.9  
>  
> # mean and standard deviation  
> mean(v)  
[1] 2.3  
> sd(v)  
[1] 1.1
```

```
> # function summary  
> v = c(1.0, 3.0, -1.5, 0, 0.5)  
> summary(v)  
Min. 1st Qu. Median Mean 3rd Qu. Max.  
-1.5 0.0 0.5 0.6 1.0 3.0  
>  
> # vector length  
> length(v)  
[1] 5  
>  
> # choose a subset  
> 1:3  
[1] 1 2 3  
> v[1:3]  
[1] 1.0 3.0 -1.5  
>  
> v1 = v[which(v>0)]  
> v1  
[1] 1.0 3.0 0.5
```

Matrices

```
> m1 = matrix(1:9, nrow=3, ncol=3)
> m1
 [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
>
> m2 = matrix(1:9, nrow=3, ncol=3, byrow=TRUE)
> m2
 [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
>
> m1 + m2
 [,1] [,2] [,3]
[1,]    2    6   10
[2,]    6   10   14
[3,]   10   14   18
```

```
>  
> # matrix dimension  
> dim(m1)  
[1] 3 3  
> dim(m2)  
[1] 3 3  
>  
> # element-wise multiplication  
> m1 * m2  
      [,1] [,2] [,3]  
[1,]     1     8    21  
[2,]     8    25    48  
[3,]    21    48    81  
>  
> # matrix multiplication  
> m1 %*% m2  
      [,1] [,2] [,3]  
[1,]   66   78   90  
[2,]   78   93  108  
[3,]   90  108  126
```

```
>  
> m1  
      [,1] [,2] [,3]  
[1,]    1    4    7  
[2,]    2    5    8  
[3,]    3    6    9  
>  
> # submatrix  
> m1[2,2]  
[1] 5  
>  
> m1[1:2,]  
      [,1] [,2] [,3]  
[1,]    1    4    7  
[2,]    2    5    8  
>  
> m1[c(1,3),2:3]  
      [,1] [,2]  
[1,]    4    7  
[2,]    6    9
```

```
>  
> m1  
      [,1] [,2] [,3]  
[1,]    1    4    7  
[2,]    2    5    8  
[3,]    3    6    9  
>  
> # function apply  
> # apply(X, MARGIN, FUN, ...)  
> # MARGIN =1 for rows, =2 for columns, =c(1,2) for rows and columns  
> # FUN is the function to be applied  
>  
> apply(m1[1:2], 2, sum)  
[1] 3 9 15  
>  
> apply(m1[c(1,3),2:3], 1, mean)  
[1] 5.5 7.5  
>  
>  
>
```

```
> # diag(1,3) creates a 3x3 diagonal matrix with 1s on the diagonal
>
> m3 = diag(1,3) + matrix(c(0,1,2,0,0,1,0,0,0),nrow=3)
> m3
      [,1] [,2] [,3]
[1,]    1    0    0
[2,]    1    1    0
[3,]    2    1    1
>
> # matrix transpose
> t(m3)
      [,1] [,2] [,3]
[1,]    1    1    2
[2,]    0    1    1
[3,]    0    0    1
>
>
>
>
>
```

```
> # matrix inverse  
> m4 = solve(m3)  
> m4  
      [,1] [,2] [,3]  
[1,]    1    0    0  
[2,]   -1    1    0  
[3,]   -1   -1    1  
  
>  
> m3 %*% m4  
      [,1] [,2] [,3]  
[1,]    1    0    0  
[2,]    0    1    0  
[3,]    0    0    1  
  
>
```

rbind / cbind

```
> # Generate a sequence
> # seq(from, to, by)
> # below could use just x = seq(1, 7, 3)
> x = seq(1, 7, by=3)
> x
[1] 1 4 7
>
> # Replicate a vector x
> # rep(x, times)
> y = rep(1, 3)
> y
[1] 1 1 1
>
> # row-wise bind
> rbind(x,y)
 [,1] [,2] [,3]
x     1     4     7
y     1     1     1
>
```

```
> # column-wise bind  
> cbind(x,y)
```

	x	y
[1,]	1	1
[2,]	4	1
[3,]	7	1

Types of variables

```
> v = 1:5
> v
[1] 1 2 3 4 5
> mode(v)
[1] "numeric"
>
> a = "Hello, World :)"
> a
[1] "Hello, World :)"
> mode(a)
[1] "character"
>
> b = v==2
> b
[1] FALSE  TRUE FALSE FALSE FALSE
> mode(b)
[1] "logical"
>
>
```

```
> # factor
>
> treatments = c("placebo", "100mg", "200mg")
>
> # sample() draws a sample with or without replacement
> csamp = sample(treatments, 6, replace=TRUE)
>
> csamp
[1] "200mg"    "placebo"   "placebo"   "100mg"    "placebo"   "200mg"
>
> # as.factor() forces its argument to be an object of class factor
> as.factor(csamp)
[1] 200mg    placebo  placebo  100mg    placebo  200mg
Levels: 100mg 200mg placebo
>
> table(csamp)
csamp
 100mg  200mg placebo
      1       2       3
```

```
> # list
> p = c("regulation of apoptosis", "response to tumor")
> g = list(gene="Tp53", process=p, expression=c(1.2,9.1))
> g
$gene
[1] "Tp53"

$process
[1] "regulation of apoptosis" "response to tumor"

$expression
[1] 1.2 9.1

> g[[1]]
[1] "Tp53"
> g[1]
$gene
[1] "Tp53"
> g$gene
[1] "Tp53"
```

```
>  
> # [ can select more than one element  
> # whereas [[ and $ can select just one  
>  
> g[1:2]  
$gene  
[1] "Tp53"  
  
$process  
[1] "regulation of apoptosis" "response to tumor"  
>  
> g[[1:2]]  
Error in g[[1:2]] : subscript out of bounds  
>
```

Data Frames

A data frame is a list with class “data.frame”. Usually, its columns are vectors of the same length. A numerical or logical vector is included as is, and a character vector is coerced to be of type factor.

```
> sym = c("BRCA1", "BRCA2", "RAS1", "APC", "Tp53")
>
> # rnorm(n, mean, sd);  defaults: mean=0, sd=1
> ep1 = round(rnorm(5,0,1),2)
> ep2 = round(rnorm(5,0,1),2)
>
> dat = data.frame(sym=sym, e1=ep1, e2=ep2)
> dat
```

	sym	e1	e2
1	BRCA1	-0.59	-1.02
2	BRCA2	-1.07	1.20
3	RAS1	-1.73	-0.59
4	APC	-1.40	0.63
5	Tp53	-1.76	-0.06

```
>  
> mode(dat)  
[1] "list"  
> dim(dat)  
[1] 5 3  
> names(dat)  
[1] "sym" "e1"  "e2"  
>  
> dat1 = cbind(sym, ep1, ep2)  
> dat1[1:2,]  
      sym      ep1      ep2  
[1,] "BRCA1" "-0.59" "-1.02"  
[2,] "BRCA2" "-1.07" "1.2"  
> dat1 = as.data.frame(dat1)  
> dat1[1:2,]  
      sym    ep1    ep2  
1 BRCA1 -0.59 -1.02  
2 BRCA2 -1.07   1.2
```

R functions, datasets, and packages

- “All R functions and datasets are stored in packages. Only when a package is loaded are its contents available.”
- By default, some standard packages (e.g., `base`, `stats`) are included in the binary distribution of R and they are loaded into the R environment automatically when one opens the R interface.
- Some recommended packages are included in the binary R distribution, but are not loaded automatically.
- Contributed packages need to be installed before one can load and use them.

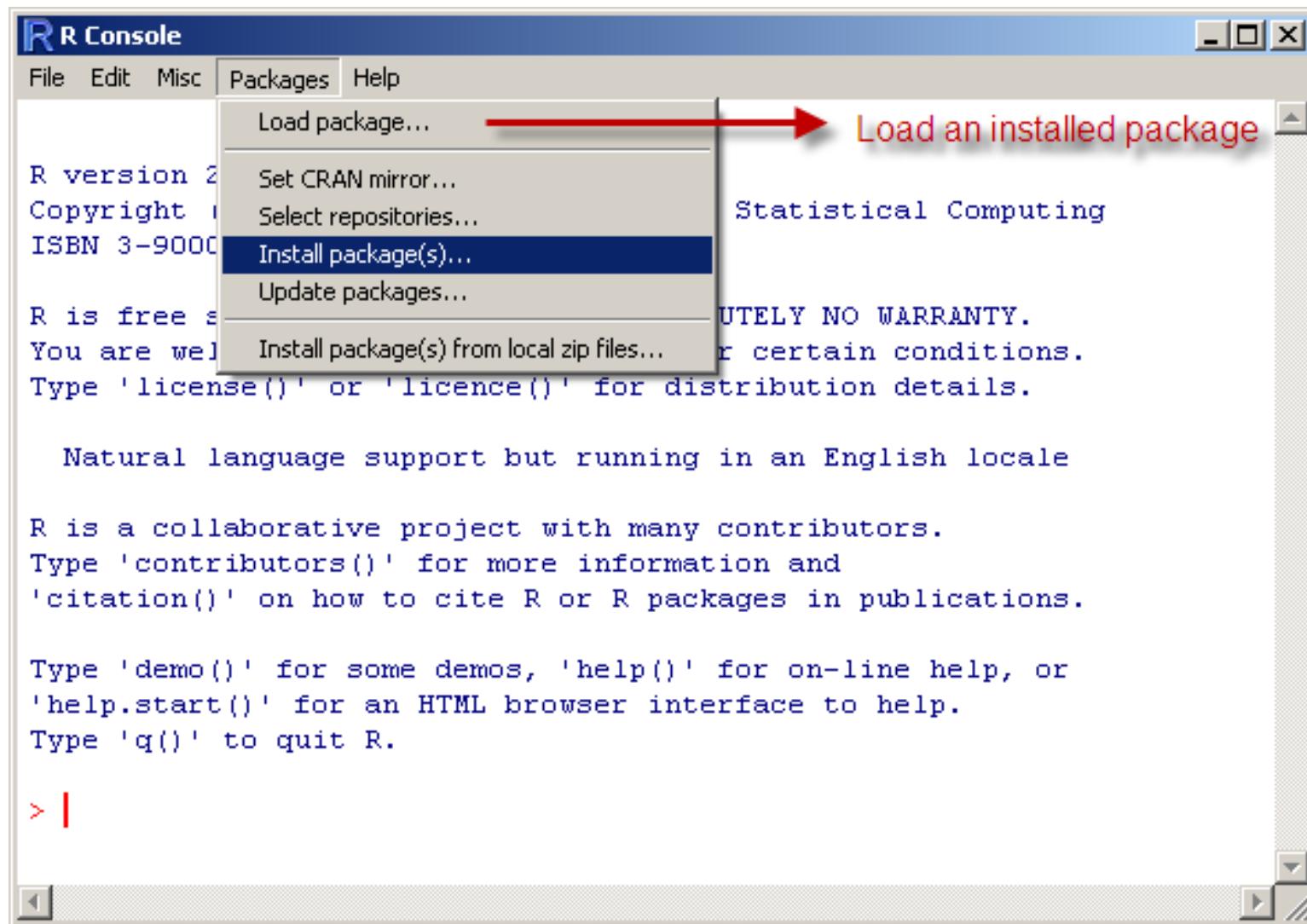


Figure 3: Load/install R packages

Help

- How does one know which function to use?
- Suppose one is looking for something related to the uniform distribution
 - `help(package="stats")`
 - `help.search("uniform")`
 - google it
- How to use a function?
 - `?runif`
 - `help(runif)`

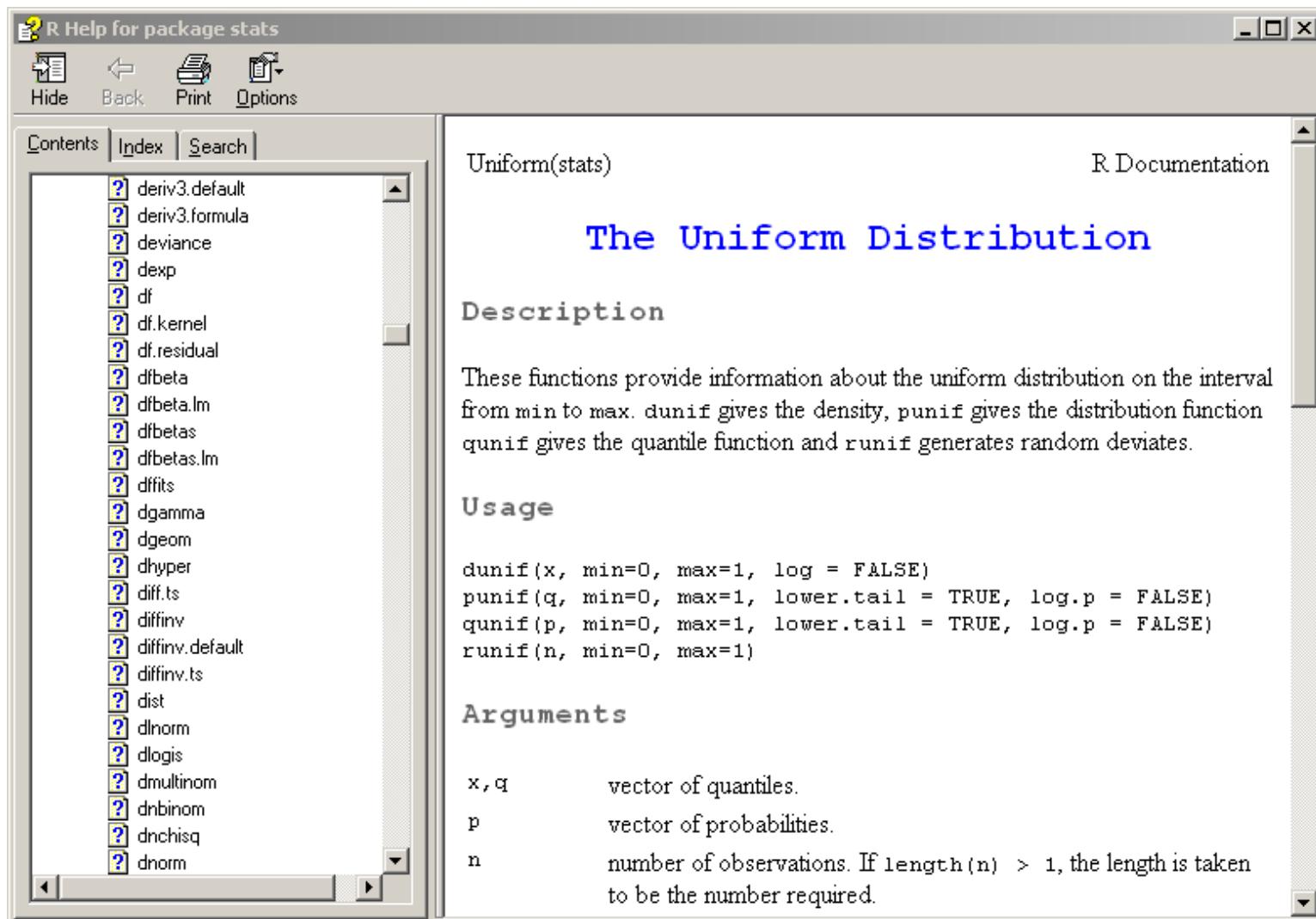


Figure 4: Help file for function “runif”

Loops and conditional execution

```
> x = runif(100)
> x[1:5]
[1] 0.1996935 0.2580256 0.2381857 0.4106282 0.8495470
> summary(x)
   Min. 1st Qu. Median      Mean 3rd Qu.      Max.
0.01639 0.32640 0.53370 0.53420 0.76960 0.99880
>
> sum(x[x>0.5 & x<0.8])
[1] 18.87469
>
> y = 0
> for(i in 1:length(x)){
+   if(x[i]>0.5 & x[i]<0.8){y = y + x[i]}
+ }
> y
[1] 18.87469
```

Writing one's own functions

```
> hi <- function(yourname, myname="David"){
+   str1 = paste("Hello, ", yourname, ", This is ", myname, sep="")
+   str1
+ }
> hi("World")
[1] "Hello, World, This is David"
> hi <- function(yourname){
+   str1 = paste("Hello,", yourname, sep=" "); substr(str1,1,8);
+ }
> hi("World")
[1] "Hello, W"
```

- Parameters, default values of the parameters
- Output is the last variable evaluated
- Different commands separated by a semi-colon, or by starting a new line

R vs. other programming languages, software packages

- S/Splus
 - R regarded as an implementation of the S language, which forms the basis of S-Plus.
 - Syntax of R is almost the same as that of S (or S-Plus).
- SAS
 - SAS procedures by default give just copies of the output, while R functions usually have all the intermediate results stored
 - One has better control in R, and R is open source
 - R has no warranty
- perl/python
 - perl has more powerful text processing facilities than R
 - There are few statistical functions in perl

R vs. c/c++

- In c/c++, need to take care of many low level things, such as memory allocation. There are some c libraries available, so one doesn't always have to start from scratch. But still, it is not always a pleasant experience spending a whole day debugging c code.
- R is interactive. This means one can track the results at each step. Gives better control, fewer bugs. R has rich libraries. Most of the time one just needs to install a library and call a function. For large scale analysis, R is slow.
- c/c++: Spend one day on programming, take 10 minutes to run the program.
- R: Spend 10 minutes writing the program. Wait 1 day for the results.

Reading data from files

```
read.table(file, header = FALSE, sep = "", quote = "\'\',",
           dec = ".", row.names, col.names,
           as.is = !stringsAsFactors,
           na.strings = "NA", colClasses = NA, nrows = -1,
           skip = 0, check.names = TRUE, fill = !blank.lines.skip,
           strip.white = FALSE, blank.lines.skip = TRUE,
           comment.char = "#",
           allowEscapes = FALSE, flush = FALSE,
           stringsAsFactors = default.stringsAsFactors(),
           encoding = "unknown")
```

Two versions of an input file

test_dat.txt

```
BRCA1 0.65 0.65
BRCA1 -1.41 -1.41
RAS1 -0.64 -0.64
APC -0.28 -0.28
Tp53 -0.27 -0.27
```

test_dat_with_header.txt

gene	sym	e1	e2
BRCA1		0.65	0.65
BRCA1		-1.41	-1.41
RAS1		-0.64	-0.64
APC'		-0.28	-0.28
Tp53		-0.27	-0.27

Reading data from files

```
> getwd()
[1] "C:/Documents and Settings/David/My Documents"
>
> setwd("G:/Z_CSCC/BIOS662_2011/From Wei Sun/R_lecture/R_lecture/dat")
> list.files()
[1] "test_dat.txt"                  "test_dat_with_header.txt"
>
> dat1 = read.table("test_dat.txt")
> dim(dat1)
[1] 5 3
> dat1[1:2,]
      V1     V2     V3
1 BRCA1  0.65  0.65
2 BRCA1 -1.41 -1.41
>
>
>
>
```

```
> dat2 = read.table("test_dat_with_header.txt",
+ header=TRUE)
Error in scan(file, what, nmax, sep, dec, quote, ...
  line 3 did not have 4 elements
In addition: Warning message:
In read.table("test_dat_with_header.txt", header = TRUE) :
  incomplete final line found by readTableHeader ...
>
> dat2 = read.table("test_dat_with_header.txt",
+ header=TRUE, sep="\t", quote = "")
>
> dim(dat2)
[1] 5 3
> dat2[1:2,]
  gene.sym    e1    e2
1   BRCA1  0.65  0.65
2   BRCA1 -1.41 -1.41
```

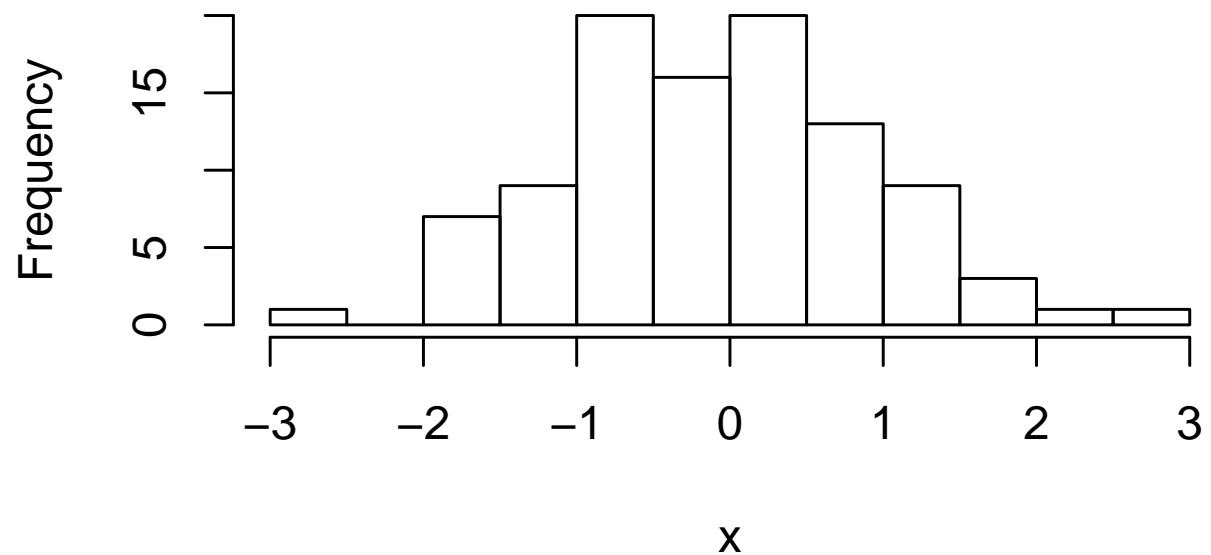
Writing data to files

```
# default  
write.table(dat2, file = "d2.txt", append = FALSE, quote = TRUE,  
            sep = " ", row.names = TRUE, col.names = TRUE)  
  
# Wei prefers these parameters  
write.table(dat, file = "d2.txt", append = FALSE, quote = FALSE,  
            sep = "\t", row.names = FALSE, col.names = TRUE)  
  
save(dat2, file = "d2.RData")  
  
load("d2.RData")
```

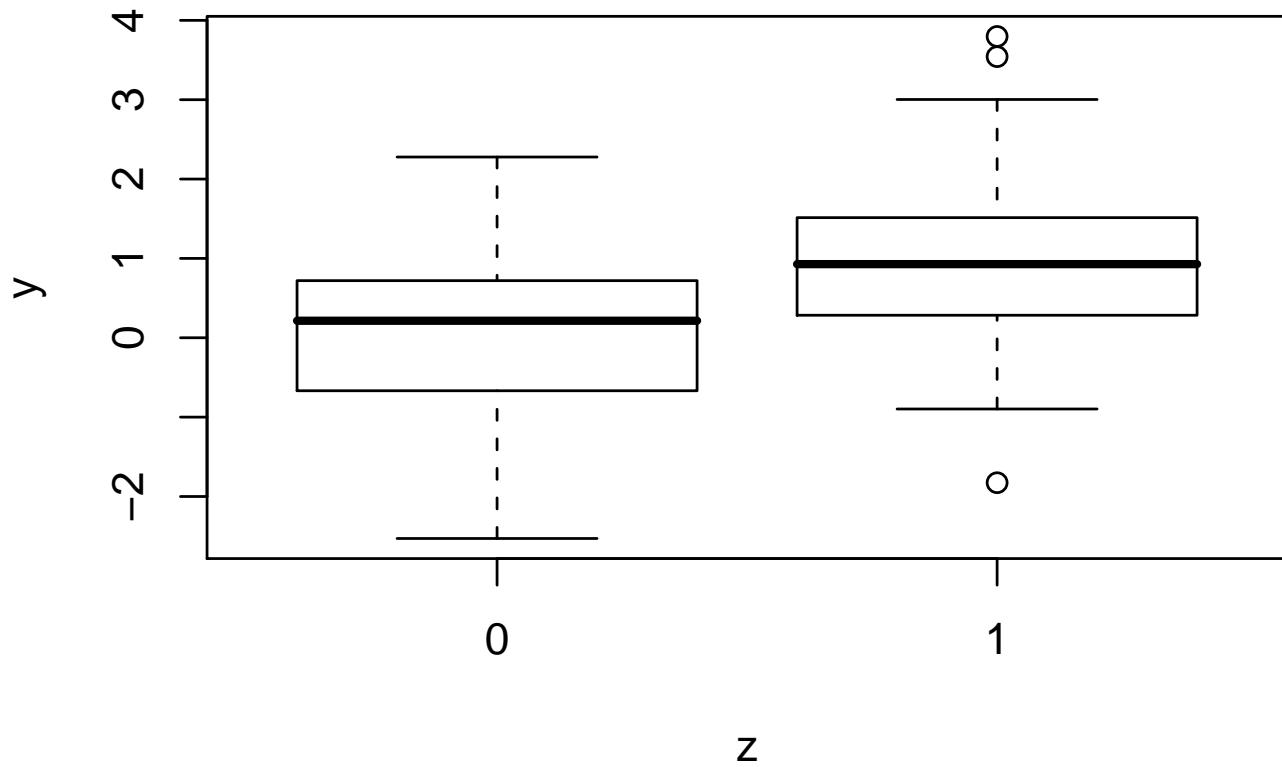
Graphics

- high-level plotting functions
 - scatter plot: `plot(x, y)`
 - histogram: `hist(x)`
 - boxplot: `boxplot(x)`
- low-level plotting commands
 - add points: `points(x, y)`
 - add lines: `lines(x, y)`, `abline(a, b)`
 - add text: `text(x, y, labels)`
 - add legend: `legend(x, y, legend)`

Histogram of x



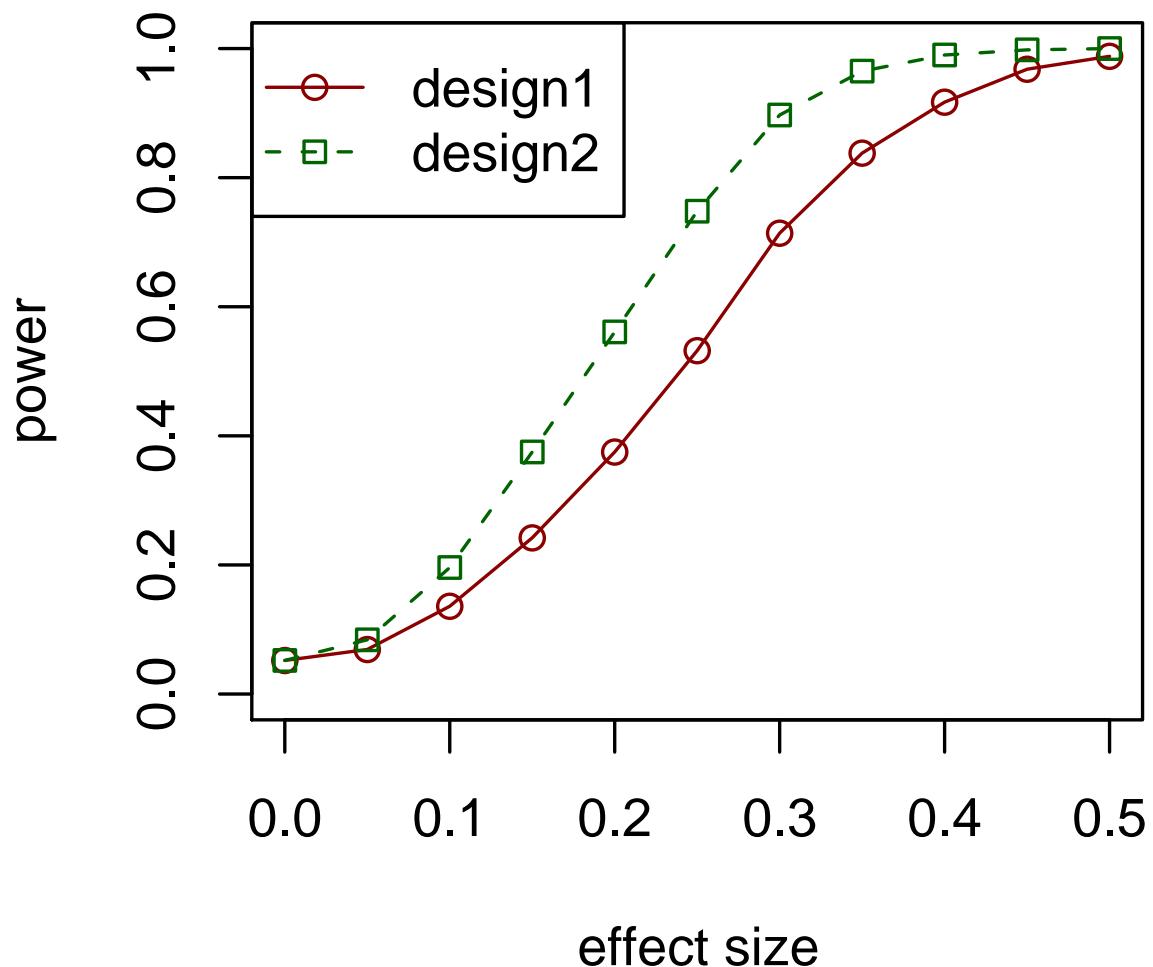
```
> x = rnorm(100)  
> hist(x)
```

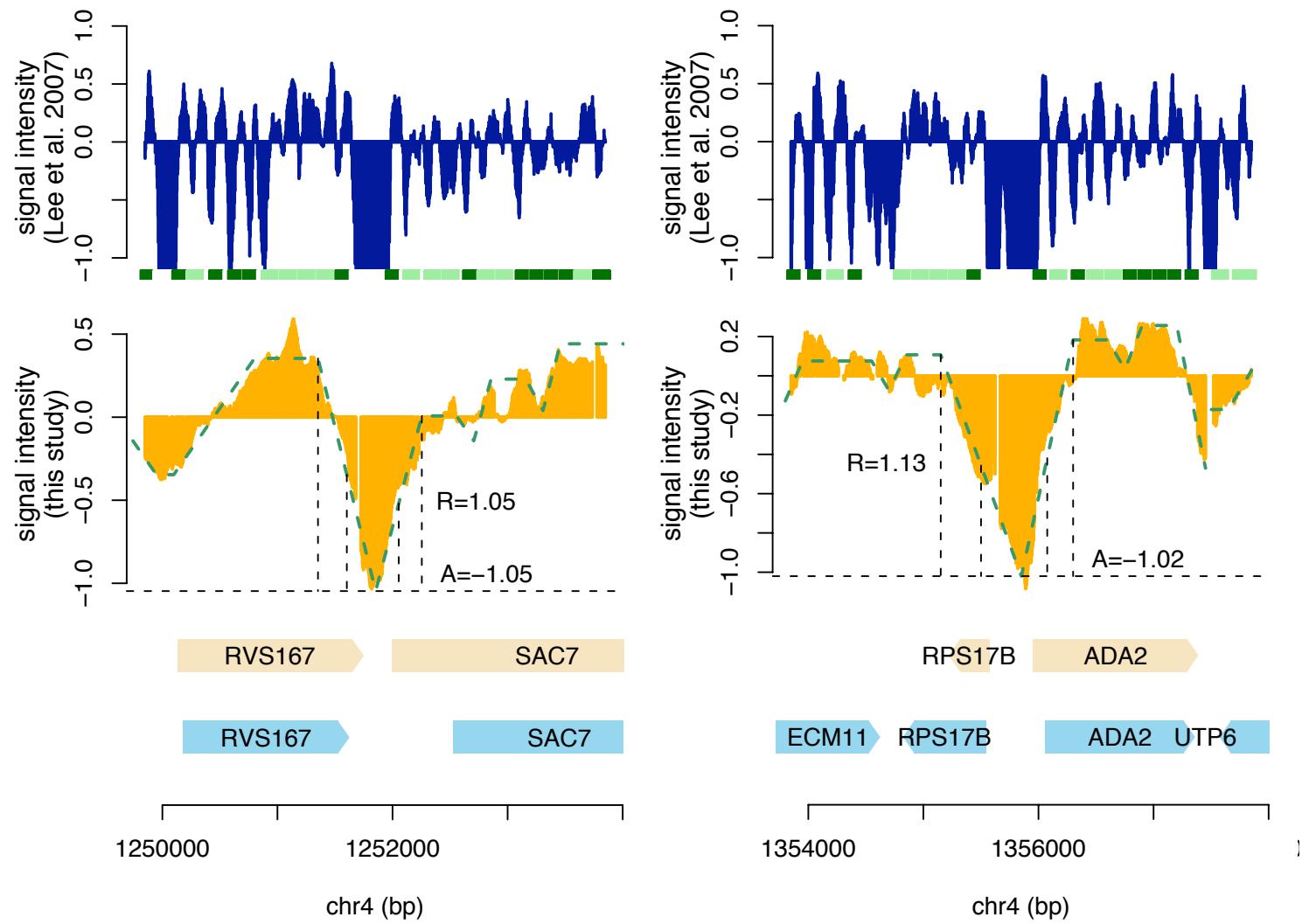


```
> y = c(rnorm(100, 0, 1), rnorm(100, 1, 1))
> z = rep(c(0,1), each=100)
> par(mar=c(5,4,1,1))
> boxplot(y~z, xlab="z", ylab="y")
```

```
> dat = read.table("power.txt", header=TRUE)
>
> dat[1:2,]
  effect design1 design2
1    0.00    0.052    0.052
2    0.05    0.069    0.084
>
> windows(width=4,height=4)
> plot(c(0, 0.5), c(0,1), type="n", ylab="power",
+   xlab="effect size", main="Power Comparison")
>
> lines(dat$effect,  dat$design1, col="darkred", lty=1)
> points(dat$effect,  dat$design1, col="darkred", pch=21)
>
> lines(dat$effect,  dat$design2, col="darkgreen", lty=2)
> points(dat$effect,  dat$design2, col="darkgreen", pch=22)
>
> legend("topleft", legend=c("design1", "design2"),
+   pch=c(21, 22), lty=c(1,2),
+   col=c("darkred", "darkgreen"))
```

Power Comparison





BIOS 662 Fall 2018

Descriptive Statistics

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Descriptive Statistics

- **Types of variables**
- Measures of location
- Measures of spread, shape
- Data displays

Types of Variables

- Definition 3.1. A *variable* is a quantity that may vary from object to object
- Definition 3.2. A *sample* or *data set* is a collection of values of one or more variables.
- Types of variables
 - Quantitative variable – intrinsically numeric
e.g. age, height, counts
 - Qualitative (categorical) – intrinsically non-numeric
e.g. gender, state, country

Types of Variables

- Qualitative (categorical) - intrinsically non-numeric
 - Binary, dichotomous
 - e.g., alive/dead, female/male
 - Ordinal - natural ordering
 - e.g., diagnosis (certain, probable, unlikely, ...)
 - e.g., attitude (strongly agree, agree, neutral, ...)
 - Nominal - no natural ordering
 - e.g., religion, race
- In recording qualitative data (and using them in analyses), numeric values may be assigned
- Some “values” may have special meaning, such as missing, N/A, unknown

Descriptive Statistics

- Types of variables
- **Measures of location**
- Measures of spread, shape
- Data displays

Definition 3.10. A *statistic* is a numerical characteristic of a sample

Measures of Location

- (Arithmetic) Mean
- Percentiles
- Median
- Mode
- Geometric mean

Arithmetic Mean

- Data:

$$x_1, x_2, \dots, x_n$$

- Mean:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Example

Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

Mean: $\bar{x} = \frac{1}{4}(5 + 10 + 6 + 11) = \frac{32}{4} = 8$

Reporting of Decimals

- Report mean with one more significant digit than the observations
- Example:

If x is measured in whole numbers and $\bar{x} = 6.345$, report $\bar{x} = 6.3$

Properties of the Mean

- Let c be any constant
- If

$$y_i = x_i + c \quad \text{for } i = 1, 2, 3, \dots, n,$$

then

$$\bar{y} = \bar{x} + c$$

- If

$$y_i = cx_i \quad \text{for } i = 1, 2, 3, \dots, n,$$

then

$$\bar{y} = c\bar{x}$$

Properties of the Mean – Example

- A sample of birth weights in a hospital found

$$\bar{y} = 3166.9 \text{ grams}$$

- 1 oz = 28.35 g
- Therefore the mean in ounces is

$$\bar{x} = \frac{\bar{y}}{28.35} = 111.7$$

Order Statistics

- Data: x_1, x_2, \dots, x_n
- Order data from smallest to largest

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

- $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ are *order statistics*

- Note

$$x_{(1)} = \min\{x_1, x_2, \dots, x_n\}$$

$$x_{(n)} = \max\{x_1, x_2, \dots, x_n\}$$

- $(1), (2), \dots, (n)$ are the *ranks* of the observations
- The textbook defines a *half rank* such that the value at the half rank $(i + 1/2)$ is $(x_{(i)} + x_{(i+1)})/2$

Example

- Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

- Order statistics:

$$x_{(1)} = 5, x_{(2)} = 6, x_{(3)} = 10, x_{(4)} = 11$$

$$x_{(2.5)} = (6 + 10)/2 = 8$$

Percentiles

- Intuitive definition: the x *percentile* is such that $x\%$ of the observations are less than that value
- Also known as sample *quantile*

Percentiles: Text Definition

- The $(p \times 100)^{\text{th}}$ percentile of a sample of size n is

$$\hat{\xi}_p = \begin{cases} y_{(np+p)} & \text{if } np + p \text{ is an integer} \\ \{y_{(\lfloor np+p \rfloor)} + y_{(\lceil np+p \rceil)}\}/2 & \text{otherwise} \end{cases}$$

for $0 < p < 1$

- Note:
 - $\lfloor y \rfloor$ is the greatest integer $\leq y$; (the *floor* function)
 - $\lceil y \rceil$ is the smallest integer $\geq y$; (the *ceiling* function)
- Compare with Definition 3.11 of text: $P^{\text{th}} \text{ percentile}$ is the value with rank $(P/100)(1+n)$. If this rank is not an integer, it is rounded to the nearest half rank.

Percentiles: General Form

- General form (Hyndman and Fan, *Am Stat* 1996)

$$\hat{\zeta}_p = (1 - \gamma)y_{(j)} + \gamma y_{(j+1)}$$

where $j = \lfloor pn + m \rfloor$ for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$.

- Let $g = pn + m - j$
- If $m = p$ then $j = \lfloor pn + p \rfloor$ and we set

$$\gamma = \begin{cases} 0 & \text{if } g = 0 \\ 1/2 & \text{if } g > 0 \end{cases}$$

we recover the text definition

Percentiles: Software

- SAS PROC UNIVARIATE: 5 definitions of percentile
- R: 9 definitions

R “quantile()” Function

```
> ?quantile
quantile                  package:stats          R Documentation
```

Sample Quantiles

Description:

The generic function 'quantile' produces sample quantiles corresponding to the given probabilities. The smallest observation corresponds to a probability of 0 and the largest to a probability of 1.

Usage:

```
quantile(x, ...)

## Default S3 method:
quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
         names = TRUE, type = 7, ...)
```

Arguments:

x: numeric vectors whose sample quantiles are wanted.

probs: numeric vector of probabilities with values in [0,1].

na.rm: logical; if true, any 'NA' and 'NaN''s are removed from 'x' before the quantiles are computed.

names: logical; if true, the result has a 'names' attribute. Set to 'FALSE' for speedup with many 'probs'.

type: an integer between 1 and 9 selecting one of the nine quantile algorithms detailed below to be used.

....: further arguments passed to or from other methods.

Types:

'quantile' returns estimates of underlying distribution quantiles based on one or two order statistics from the supplied elements in 'x' at probabilities in 'probs'. One of the nine quantile algorithms discussed in Hyndman and Fan (1996), selected by 'type', is employed.

Percentiles: Class Definition

- The $(p \times 100)^{\text{th}}$ percentile of a sample:

$$\hat{\xi}_p = \begin{cases} y_{(\lfloor np \rfloor + 1)} & \text{if } np \text{ is not an integer} \\ \{y_{(np)} + y_{(np+1)}\}/2 & \text{if } np \text{ is an integer} \end{cases}$$

for $0 < p < 1$

- Definition 2 of R / Hyndman and Fan: $m = 0$, so $j = \lfloor pn \rfloor$, $g = pn - j$, and

$$\gamma = \begin{cases} 1 & \text{if } g > 0 \\ 1/2 & \text{if } g = 0 \end{cases}$$

- Definition 5 of SAS

Example

- Suppose $n = 278$ and we want the 75th percentile

$$np = 278 \times 0.75 = 208.5$$

so

$$\hat{\zeta}_{0.75} = x_{(209)}$$

- R

```
> x <- 1:278
> quantile(x,0.75,type=2)
75%
209
```

Example: SAS

```
data;  
    infile "C:\BIOS662\2018fall\percentile.txt";  
    input x;  
  
proc univariate; var x; run;
```

The UNIVARIATE Procedure

Variable: x

Quantiles (Definition 5)

Quantile	Estimate
75% Q3	209.0
50% Median	139.5
25% Q1	70.0
10%	28.0
5%	14.0
1%	3.0
0% Min	1.0

Median

- The sample median is the 50th percentile

$$\hat{\zeta}_{0.5} = \begin{cases} y_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ \{y_{(n/2)} + y_{(n/2+1)}\}/2 & \text{if } n \text{ is even} \end{cases}$$

for $0 < p < 1$

- Example

- Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

- Median:

$$\hat{\zeta}_{0.5} = \{x_{(2)} + x_{(3)}\}/2 = (6 + 10)/2 = 8$$

Mode

- The mode is the most frequently occurring value in the data set
- Example:

If $x_1 = 5$, $x_2 = 11$, $x_3 = 6$, $x_4 = 11$

then the mode is 11

Geometric Mean

- Data: $x_1, x_2, \dots, x_n > 0$
- The geometric mean of x is

$$\bar{x}_g = (x_1 \cdot x_2 \cdots x_n)^{1/n}$$

- Let $y_i = \log(x_i)$ for $i = 1, 2, \dots, n$. Then

$$\bar{x}_g = \exp(\bar{y})$$

- \bar{x}_g is used when data are of the form c^k
- Example:

Suppose $x_1 = 10$ and $x_2 = 0.1$

Then $\bar{x}_g = 1$

Comments

- The mean is the most often used measure
- The median is better if there are influential observations (it is more robust to extreme values)
- The mode is rarely used (exception: nominal data)

Example

- Duration of hospital stay in days:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 11$$

$$\hat{\zeta}_{0.5} = \bar{x} = 8, \quad \bar{x}_g = 7.6$$

- Alter last observation:

$$x_1 = 5, x_2 = 10, x_3 = 6, x_4 = 50$$

$$\hat{\zeta}_{0.5} = 8, \quad \bar{x} = 17.7, \quad \bar{x}_g = 11.1$$

Descriptive Statistics

- Types of variables
- Measures of location
- **Measures of spread, shape**
- Data Displays

Measures of Spread, Shape

- Range
- Variance and standard deviation
- Interquartile range
- Skewness, kurtosis

Range

- Range:

$$r_a = x_{(n)} - x_{(1)}$$

- Easy to calculate
- Sensitive to unusual observations (outliers)
- Usually, the larger n is, the larger r_a

Sample Variance and Standard Deviation

- Want to measure deviation from mean
- Sample variance

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

- Sample standard deviation

$$s = \sqrt{s^2}$$

Sample Variance and Standard Deviation

- An alternative form of the sample variance is

$$s_1^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Can show s^2 is unbiased for population variance σ^2 , however

$$E(s_1^2) = \sigma^2 - \frac{\sigma^2}{n}$$

- van Belle *et al.* argue for s^2 based on degrees of freedom (d.f.) (Note 3.5)

Sample Standard Deviation

- The units of s are the same as the units of x_i
- If s is large, the data are spread over a wide range
- The textbook recommends reporting the standard deviation with two more significant digits than the original observations

Properties of the Standard Deviation

- If c is a constant and

$$y_i = x_i + c,$$

then

$$s_y = s_x$$

- If

$$y_i = cx_i$$

then

$$s_y = cs_x$$

Some Approximations

- The interval $\bar{x} \pm s$ usually contains approximately 68% of the observations
- The interval $\bar{x} \pm 2s$ usually contains approximately 95% of the observations
- Approximate s by

$$s \approx \frac{\hat{\zeta}_{0.75} - \hat{\zeta}_{0.25}}{1.35}$$

- Note

$$\hat{\zeta}_{0.75} - \hat{\zeta}_{0.25}$$

is called the *interquartile range*

Comments

- $\hat{\zeta}_{0.25}$ is the lower quartile
- $\hat{\zeta}_{0.75}$ is the upper quartile
- Epidemiologists often use quantiles to categorize a continuous variable
- Splitting a variable at $\hat{\zeta}_{0.25}, \hat{\zeta}_{0.5}, \hat{\zeta}_{0.75}$ yields four (roughly) equally-sized groups
- Statisticians use quartile to refer to one of the cut-points
- Epidemiologists usually use it to refer to the categories
- Tertiles (for three groups) or quintiles (for five groups) are also often used

Symmetry and Skewness

- Informally, define *symmetry* to indicate having a uniform or even distribution about the mean
- If a distribution is symmetric,
$$\text{mean} = \text{median}$$
- Data that are not symmetric are said to be *skewed*
- *Skewness* is a measure of the degree to which a data set is skewed

Skewness

- Define r th sample moment about the mean

$$m_r = \frac{\sum_i (y_i - \bar{y})^r}{n} \text{ for } r = 1, 2, 3, \dots$$

- Text definition of sample skewness:

$$a_3 = \frac{m_3}{(m_2)^{3/2}} = \frac{\sum_i (y_i - \bar{y})^3 / n}{\{\sum_i (y_i - \bar{y})^2 / n\}^{3/2}} = \sqrt{n} \frac{\sum_i (y_i - \bar{y})^3}{\{\sum_i (y_i - \bar{y})^2\}^{3/2}}$$

- Typo in text on page 51
- In SAS PROC UNIVARIATE need to use the option VARDEF=N to obtain a_3

Interpretation?

- Text:

“skewed to the right if the mean is greater than the mode”

“Values of $a_3 > 0$ indicate ... skewness to the right”

- However, for $\{0, 2, 2, 3, 4\}$

$$\bar{x} = 2.2$$

$$\text{mode} = 2$$

$$\text{skewness} = -0.37$$

Alternative Definitions

- Another definition of skewness:

$$b_3 = \frac{n\sqrt{n-1}}{n-2} \frac{\sum_i (y_i - \bar{y})^3}{\{\sum_i (y_i - \bar{y})^2\}^{3/2}}$$

- b_3 is the default in SAS
- Many more definitions; cf. Joanes and Gill (JRSS D 1998)

Kurtosis

- *Kurtosis* is a measure of the flatness or peakedness of a distribution; degree of archedness; thickness of tails
- Text definition of *sample kurtosis*:

$$a_4 = \frac{m_4}{(m_2)^2} = \frac{\sum_i (y_i - \bar{y})^4 / n}{\{\sum_i (y_i - \bar{y})^2 / n\}^2} = n \frac{\sum_i (y_i - \bar{y})^4}{\{\sum_i (y_i - \bar{y})^2\}^2}$$

- Typo in text on page 51

Kurtosis: SAS

- In SAS PROC UNIVARIATE need to use the option VARDEF=N to obtain a_4

$$b_4 = \frac{1}{n} \sum \left(\frac{y_i - \bar{y}}{s_1} \right)^4 - 3$$

i.e.,

$$b_4 = \frac{\sum (y_i - \bar{y})^4 / n}{s_1^4} - 3$$

i.e.,

$$b_4 = \frac{m_4}{(m_2)^2} - 3 = a_4 - 3$$

- Why minus 3?

Descriptive Statistics

- Types of variables
- Measures of location
- Measures of spread, shape
- **Data displays**

Data displays

- Simplest form is a line listing – essentially a tabular display with each line containing the data for a single person (or other unit of observation)
- A *frequency table* gives the frequency of observations within a set of ordered intervals
 - Intervals should be mutually exclusive and exhaustive
 - 8 to 10 intervals is usually sufficient
 - With the exception of the end intervals, the length of the intervals should be constant

Frequency Table – Example: Table 3.6

Blood Pressure	Native Japanese	Generation	
		1st	2nd
≤ 104	218	4	23
106-114	272	23	132
116-124	337	49	290
126-134	362	33	347
136-144	302	41	346
146-154	261	38	202
156-164	166	23	109
> 164	314	52	112
Total	2232	263	1561

Frequency Tables

- Table on previous slide an example of an *empirical frequency distribution*
- Difficult to compare blood pressure distributions because of different sample sizes
- Divide by sample size to get *empirical relative frequency distribution*

ERFD – Example: Table 3.7

Blood Pressure	Native Japanese	Generation	
		1st	2nd
≤ 104	0.098	0.015	0.015
106-114	0.122	0.087	0.085
116-124	0.151	0.186	0.186
126-134	0.162	0.125	0.222
136-144	0.135	0.156	0.222
146-154	0.117	0.144	0.129
156-164	0.074	0.087	0.070
> 164	0.141	0.198	0.072
Total	2232	263	1561

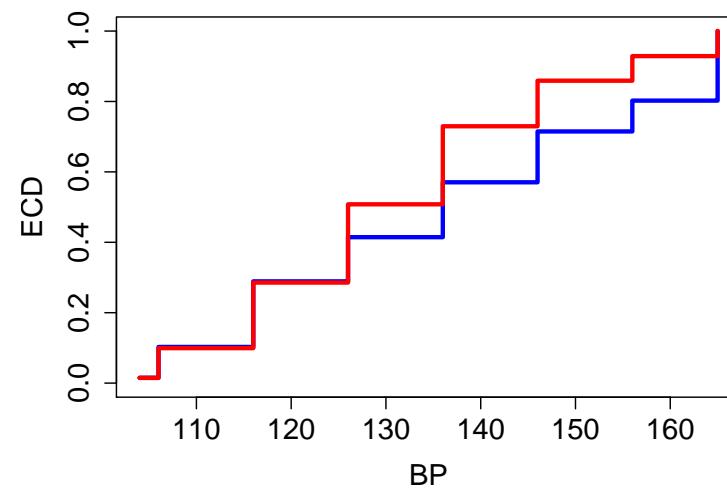
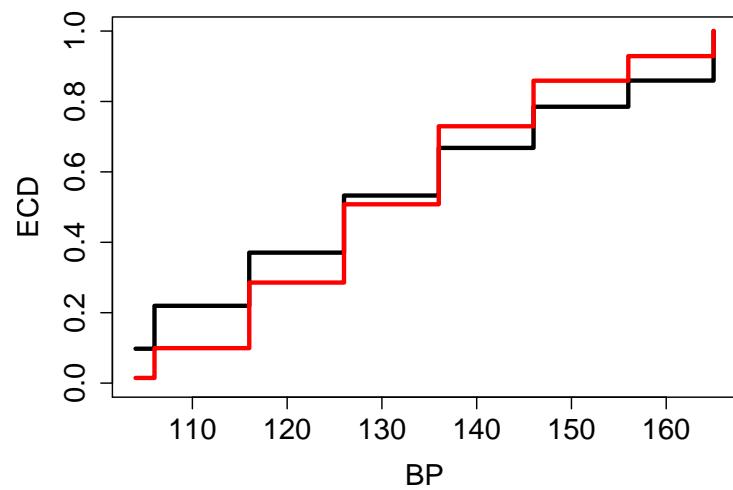
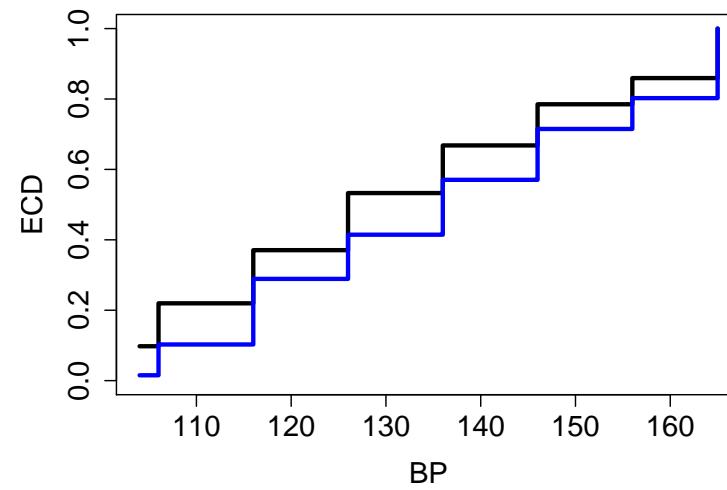
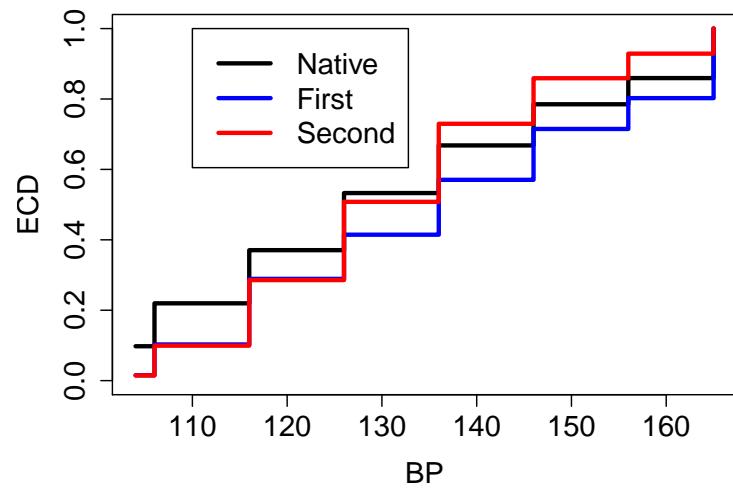
Empirical Distribution Function

- Definition 3.9. The *empirical cumulative distribution* of a variable is a listing of the values of the variable with the proportion of observations less than or equal to each value (cumulative proportion)
- Also known as the *empirical distribution function* (EDF)
- Does not necessarily entail binning (that is, grouping into intervals)

ECD – Example

Blood Pressure	Native Japanese	Generation	
		1st	2nd
≤ 104	0.098	0.015	0.015
≤ 114	0.220	0.103	0.100
≤ 124	0.371	0.289	0.285
≤ 134	0.533	0.414	0.507
≤ 144	0.668	0.570	0.729
≤ 154	0.785	0.715	0.858
≤ 164	0.859	0.802	0.928
$< \infty$	1.000	1.000	1.000
Total	2232	263	1561

ECD – Example



Graphs

- ECD/EDF
- Histogram
- Stem and leaf plot
- Box plot
- Trellis/conditional plots

Histogram

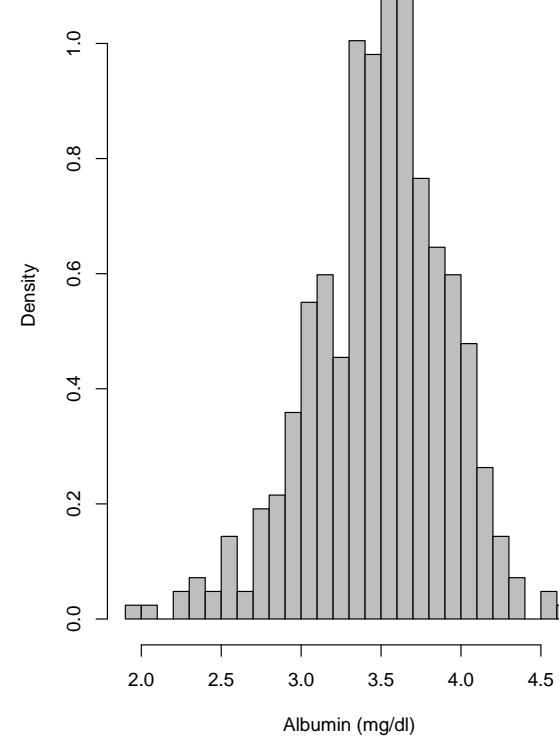
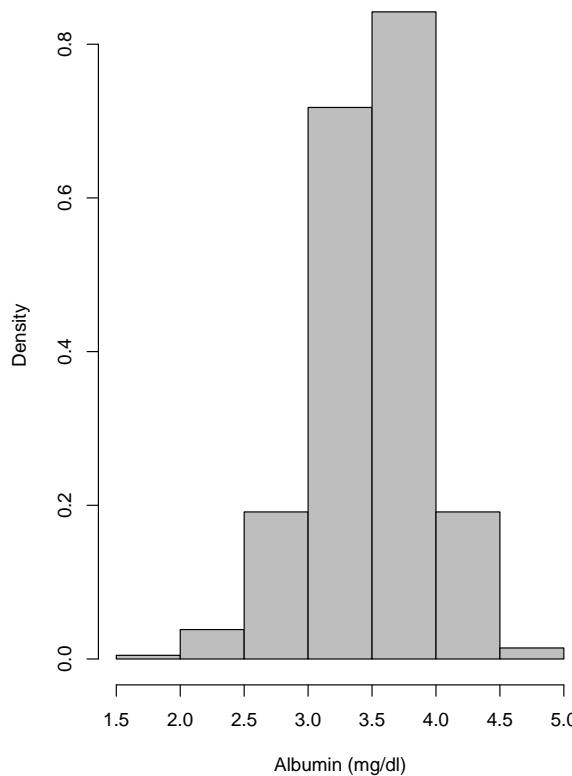
- Data are divided into intervals as in a frequency table
- A histogram is a bar graph with the area of each bar equal to the relative frequency in the interval.
- Can compare histograms from samples of different size
- Intervals need not be the same width
- Consider effect of choice of interval width (Figure 3.1 in Text)

Histogram: Example (Figure 3.1 in text)

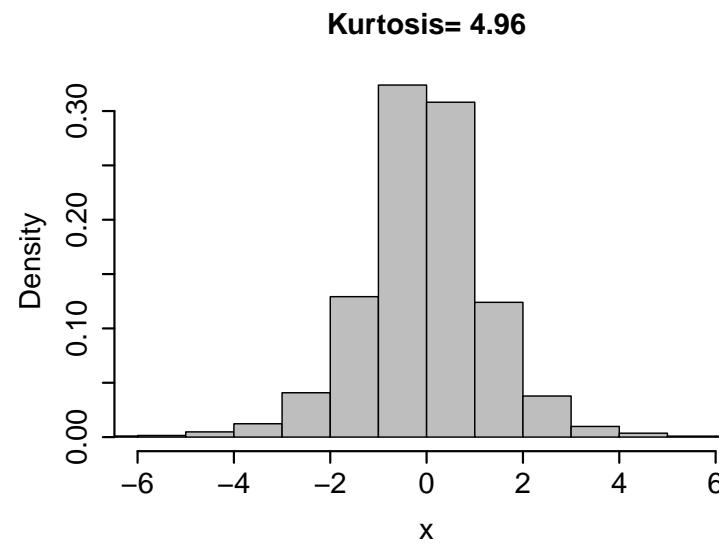
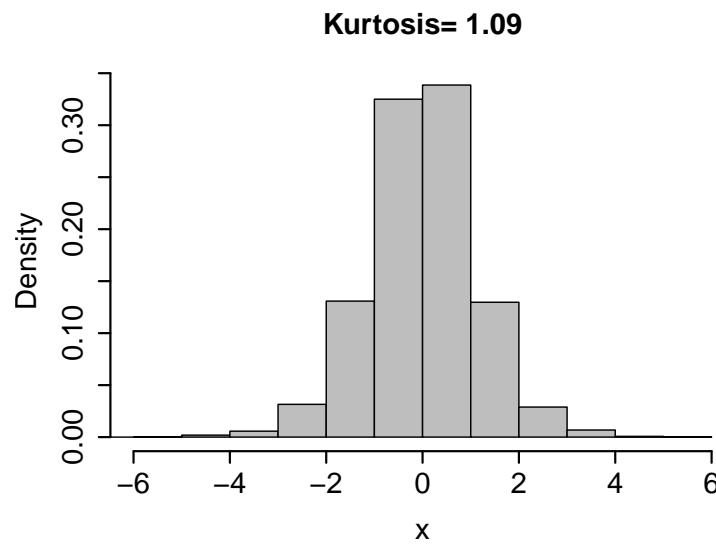
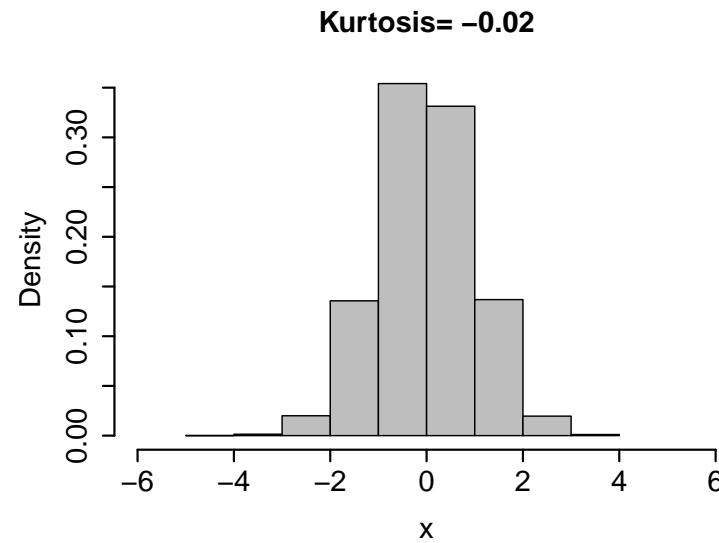
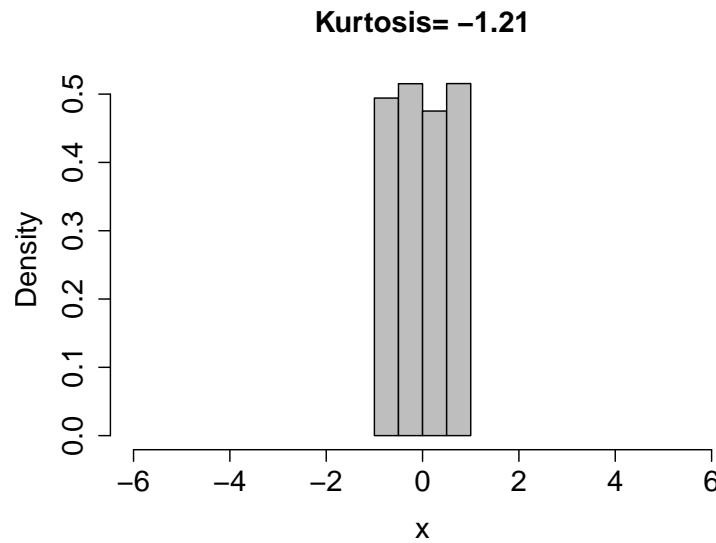
```
> par(mfcol=c(1,2))

> hist(liver$albumin,col="gray",xlab="Albumin (mg/dl)",breaks=7,freq=F,main="")

> hist(liver$albumin,col="gray",xlab="Albumin (mg/dl)",breaks=30,freq=F,main="")
```



Histograms with Various Values for Kurtosis



Stem and Leaf Plot

- Stem consists of leading digits
- Leaves consist of last digit
- Example: for $x = 496$, stem = 49, leaf = 6
- Make a column of stems from smallest to largest
- To the right of each stem, list in a row the leaves, in ascending order.
- Note: there will be one leaf for each observation

Stem and Leaf Plot: Example

```
> stem(liver$albumin)
```

The decimal point is 1 digit(s) to the left of the |

```
18 | 6
20 | 0
22 | 37138
24 | 3834468
26 | 048345557
28 | 00123447990333445666778
      00112223445556788999
30 | 0000111122333445666777880011222234455567889999
32 | 0000012233344566666999011123344444555555566666777789
34 | 0000000111222233333444556666788888999000000001112222333444455555+5
36 | 0000000011112223333334444555555566667778889999900000002233344445556+2
38 | 0000011122333333455555567799900012233334445567788889999
40 | 00111334467888889999003456678999
42 | 022340088
44 | 022
46 | 4
```

Stem and Leaf Plot: Example

```
> stem(liver$albumin, width=100)
```

The decimal point is 1 digit(s) to the left of the |

```
18 | 6
20 | 0
22 | 37138
24 | 3834468
26 | 048345557
28 | 001234479903334456667778
      0011222234455567889999
30 | 0000111122333445666777880011222234455567889999
32 | 00000122333445666669990111123344444555555566666777789
34 | 0000000111222223333444556666788888999000000001111222233344445555666666777778889
36 | 00000000111122233333444455555556666777788899990000000223334444555666666777778999
38 | 0000011122333333455555567799900012233334445567788889999
40 | 00111334467888889999003456678999
42 | 022340088
44 | 022
46 | 4
```

```
> stem(liver$albumin, scale=2) # scale changes the (vertical) length of the display
```

The decimal point is 1 digit(s) to the left of the |

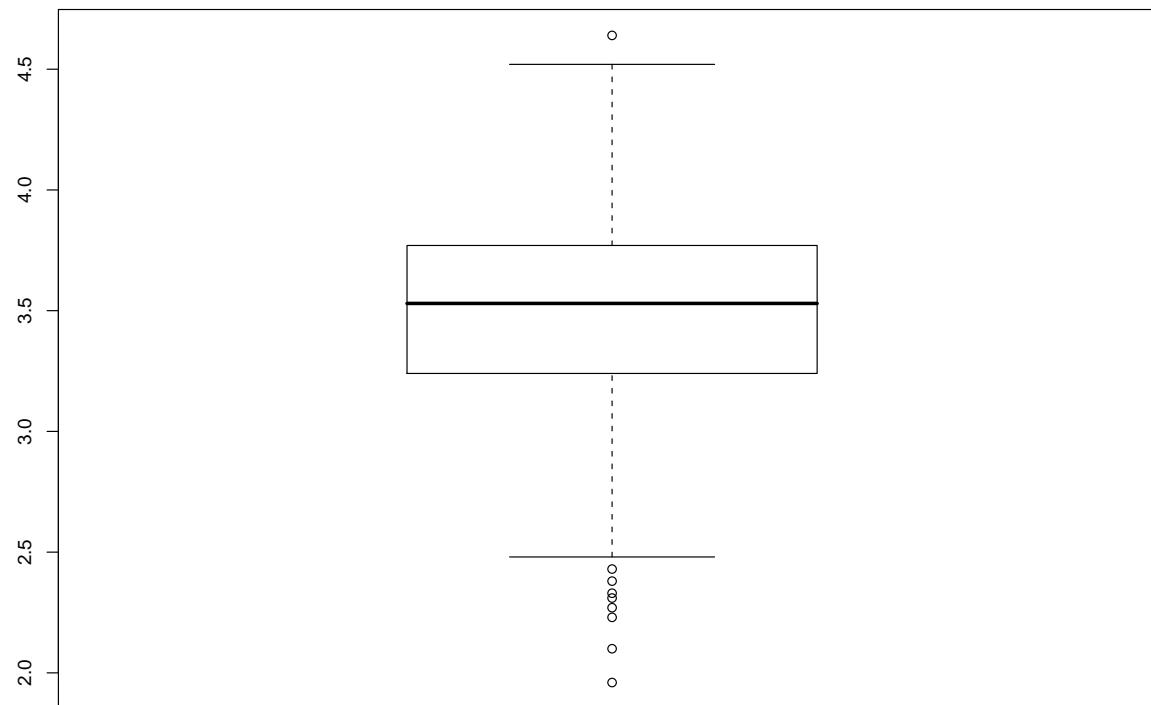
```
19 | 6
20 |
21 | 0
22 | 37
23 | 138
24 | 38
25 | 34468
26 | 048
27 | 345557
28 | 0012344799
29 | 033344566778
30 | 000011122333445666777888
31 | 0011222234455567889999
32 | 00000122333445666666999
33 | 011112334444455555555566666777789
34 | 00000001112222233333444556666788888999
35 | 000000001111222233344445555566666777778889
36 | 0000000011112222333334444555555666677778889999
37 | 00000002233344445556666677778999
38 | 0000011122333334555555677999
39 | 00012233334445567788889999
40 | 00111334467888889999
41 | 003456678999
42 | 02234
43 | 0088
44 | 0
45 | 22
46 | 4
```

Box Plot

- The top of the box is the 75th percentile ($\hat{\zeta}_{0.75}$);
the bottom is the 25th percentile ($\hat{\zeta}_{0.25}$)
- A line through the box is drawn at the median
- The lines extending out of the box (*whiskers*) may extend to
 - the 90th and 10th percentiles
 - the largest and smallest values
 - largest observation $\leq \hat{\zeta}_{0.75} + 1.5 \times \text{IQR}$;
smallest observation $\geq \hat{\zeta}_{0.25} - 1.5 \times \text{IQR}$
(Text is wrong! cf. Tukey 1977, Chambers *et al.* 1983)
- Data beyond whiskers may be plotted individually

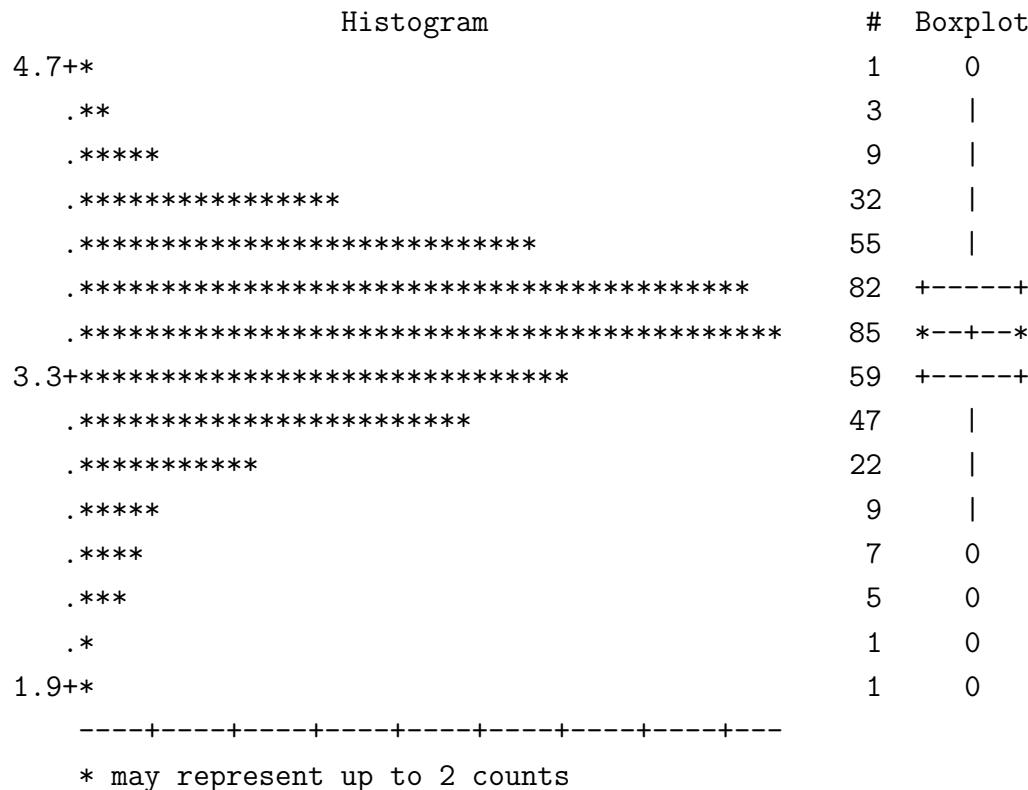
Box Plot: Example Using R

```
> boxplot(liver$albumin)
```



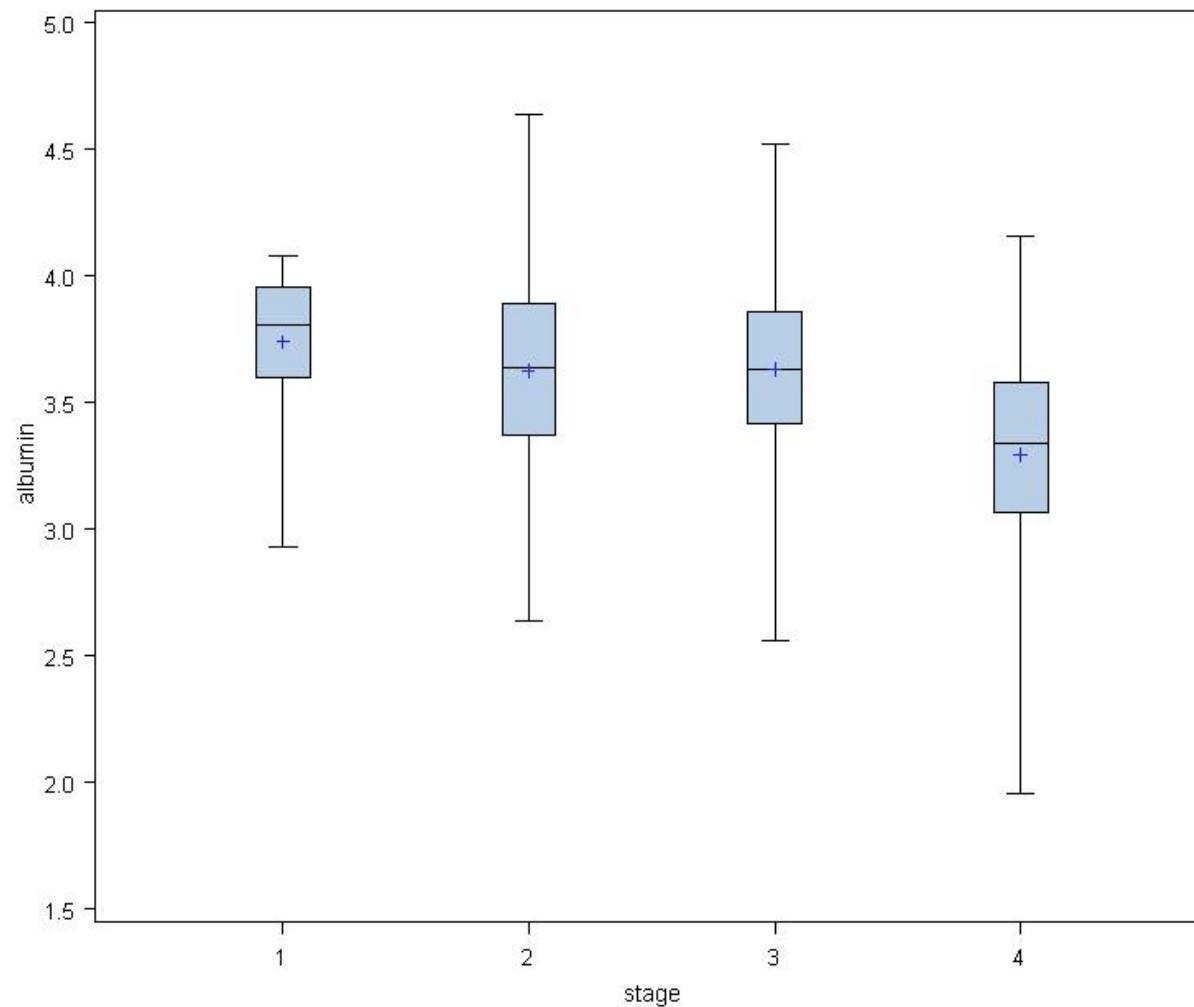
Histogram and Box Plot: Example Using SAS

```
proc univariate plot;  
  var albumin;
```



Box Plot: Example Using SAS

```
proc boxplot;  
  plot albumin*stage;
```



Box Plot

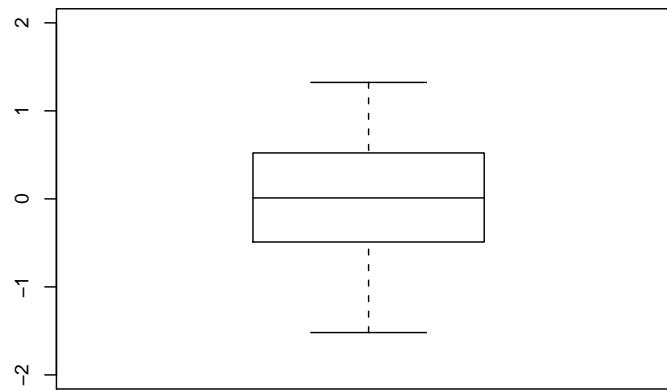
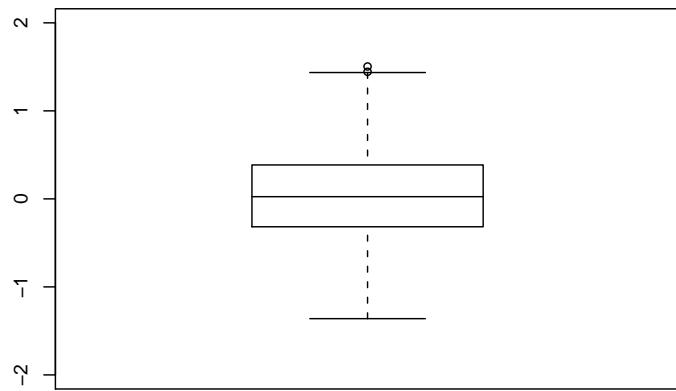
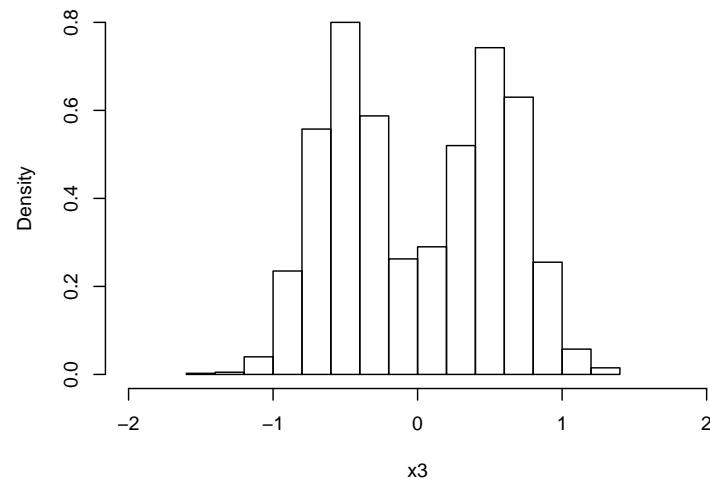
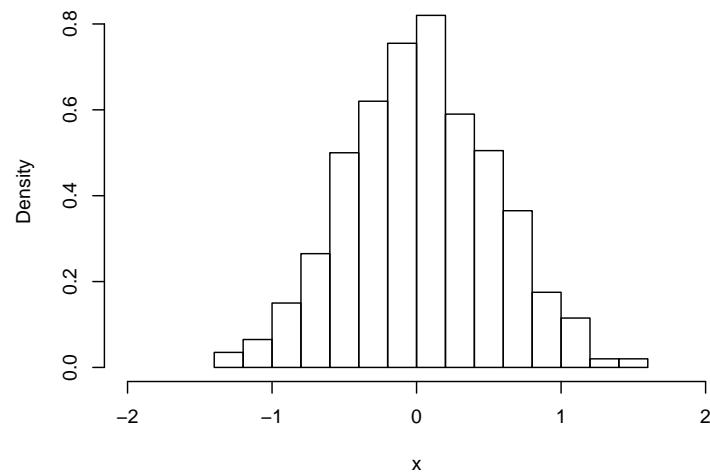
- What proportion of the data should we expect to be between the whiskers?
- If data normally distributed,
 - 95-98% for $6 \leq n \leq 20$,
 - 99% for $n > 20$
 - Ref: Hoaglin *et al.* (JASA 1986)
- Note

$$1.5 \times IQR \approx 1.5(1.35)s \approx 2s$$

so whiskers cover

$$\approx \hat{\zeta}_{0.5} \pm 2.68s$$

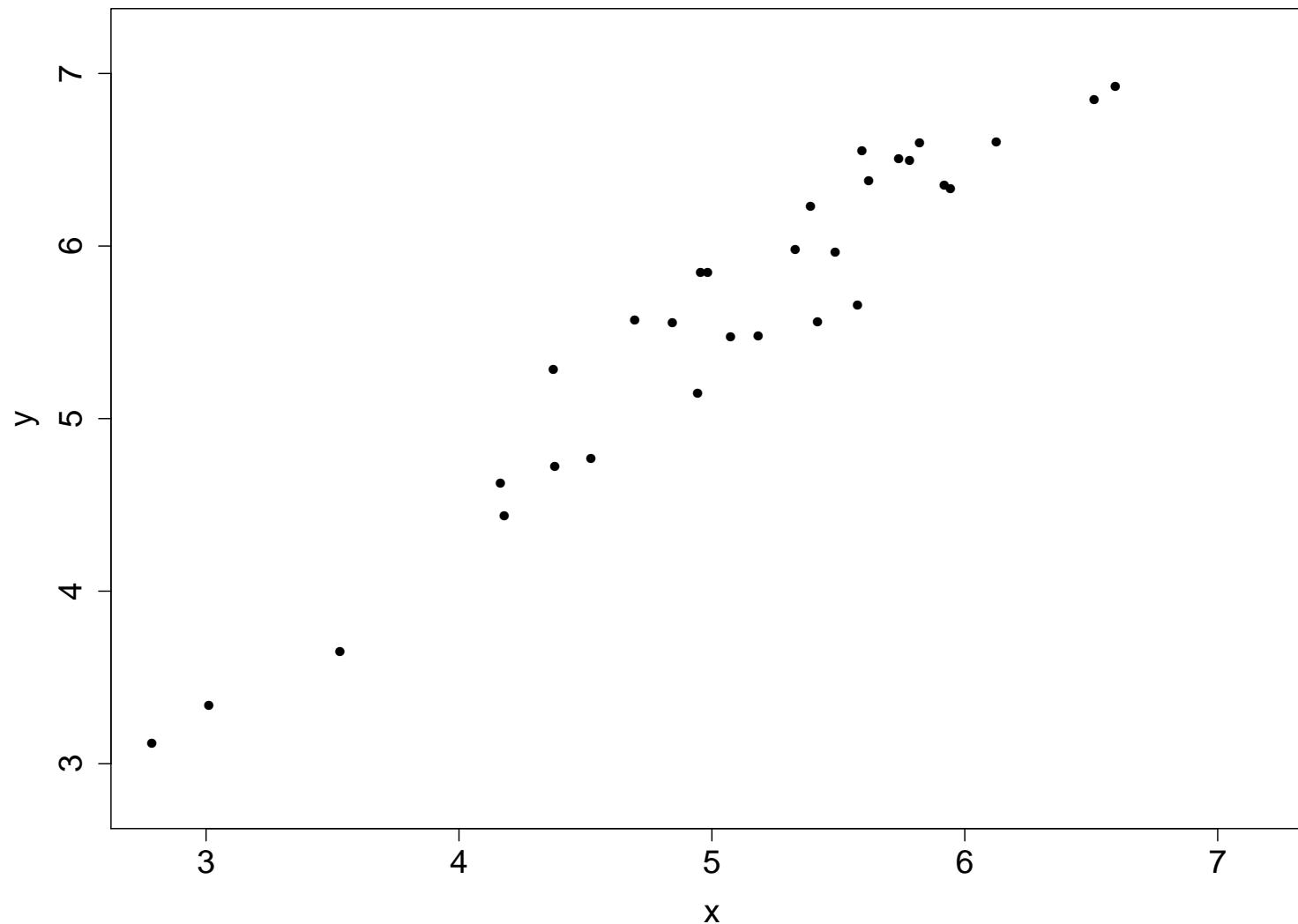
Box Plot and Histogram: Example



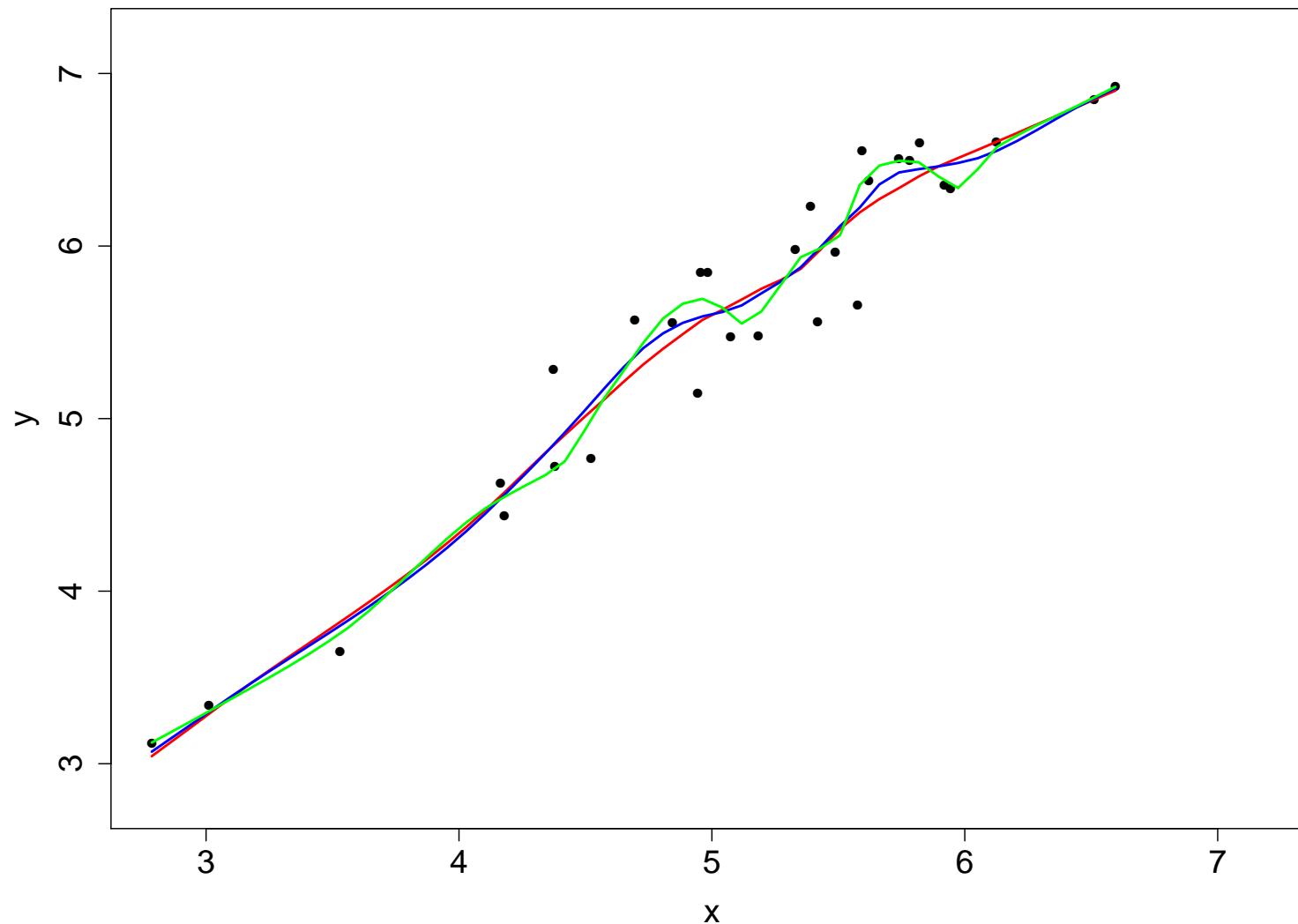
Multivariate Plots

- Describe relationships/associations between more than one variable
- Scatterplots
 - Simple for two variables
 - Add color, symbols for > 2 variables
- Trellis/conditional plots

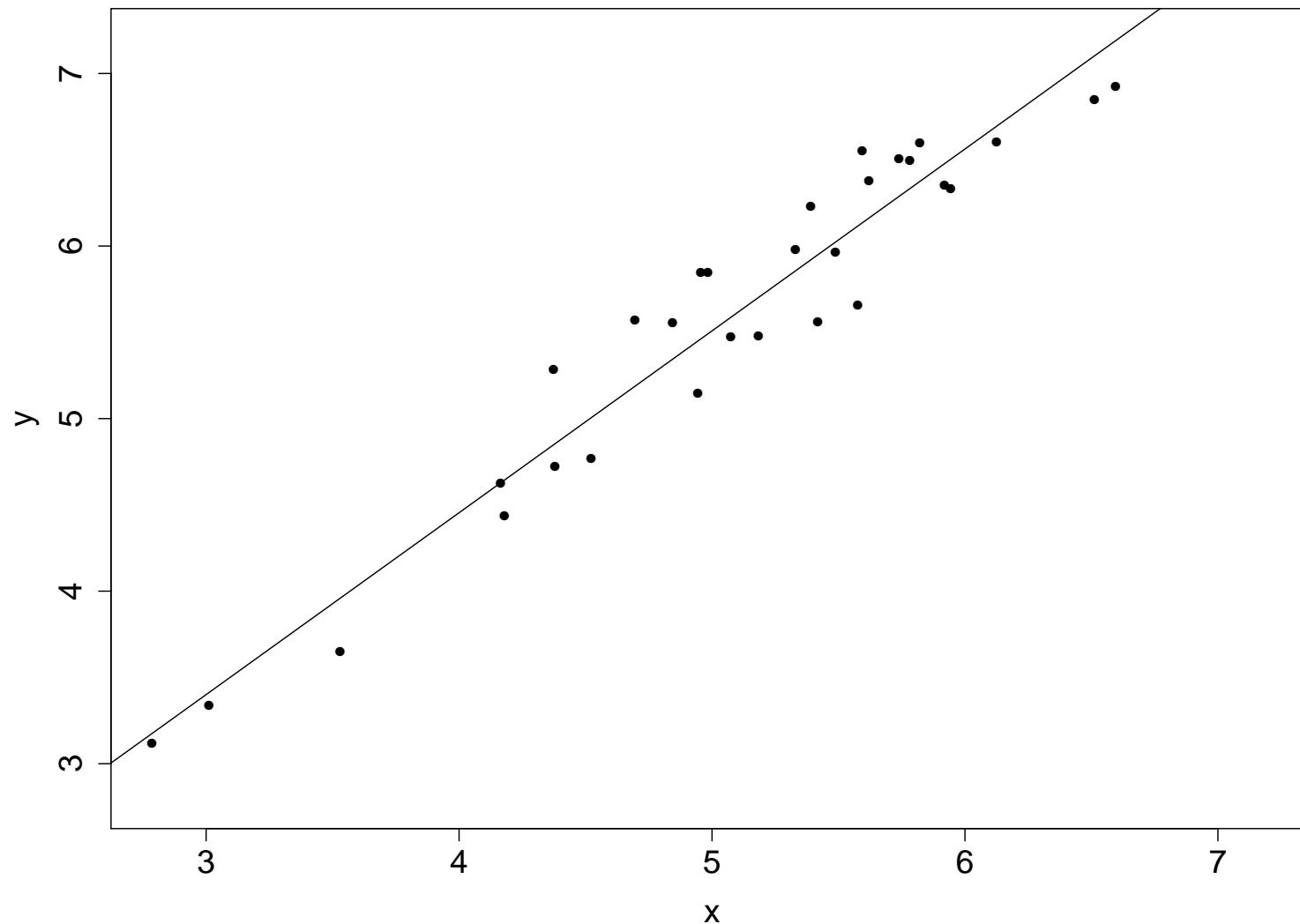
Scatterplot: Example



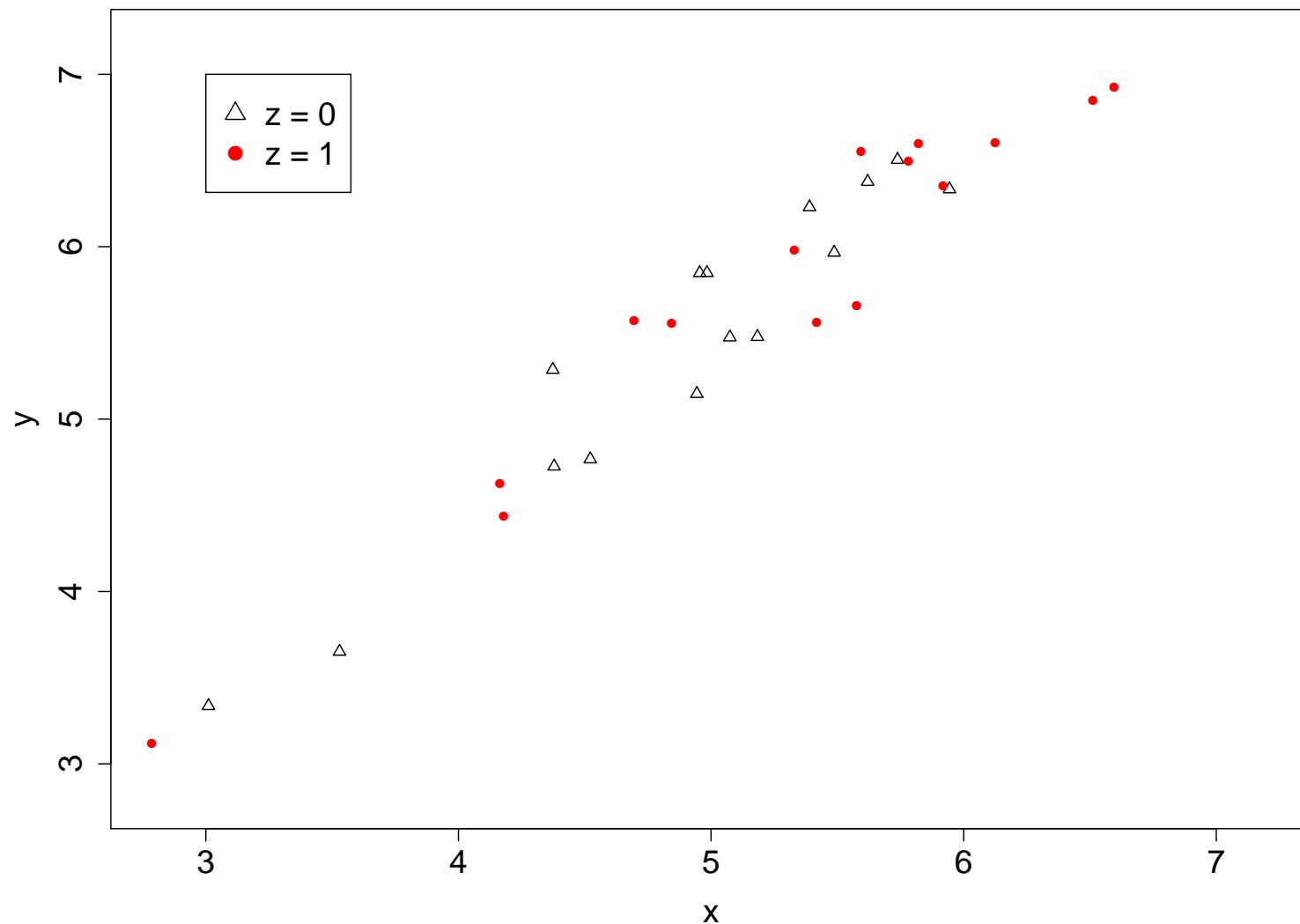
Scatterplot Example cont.



Scatterplot Example cont.



Scatterplot Example cont.



Trellis Plots

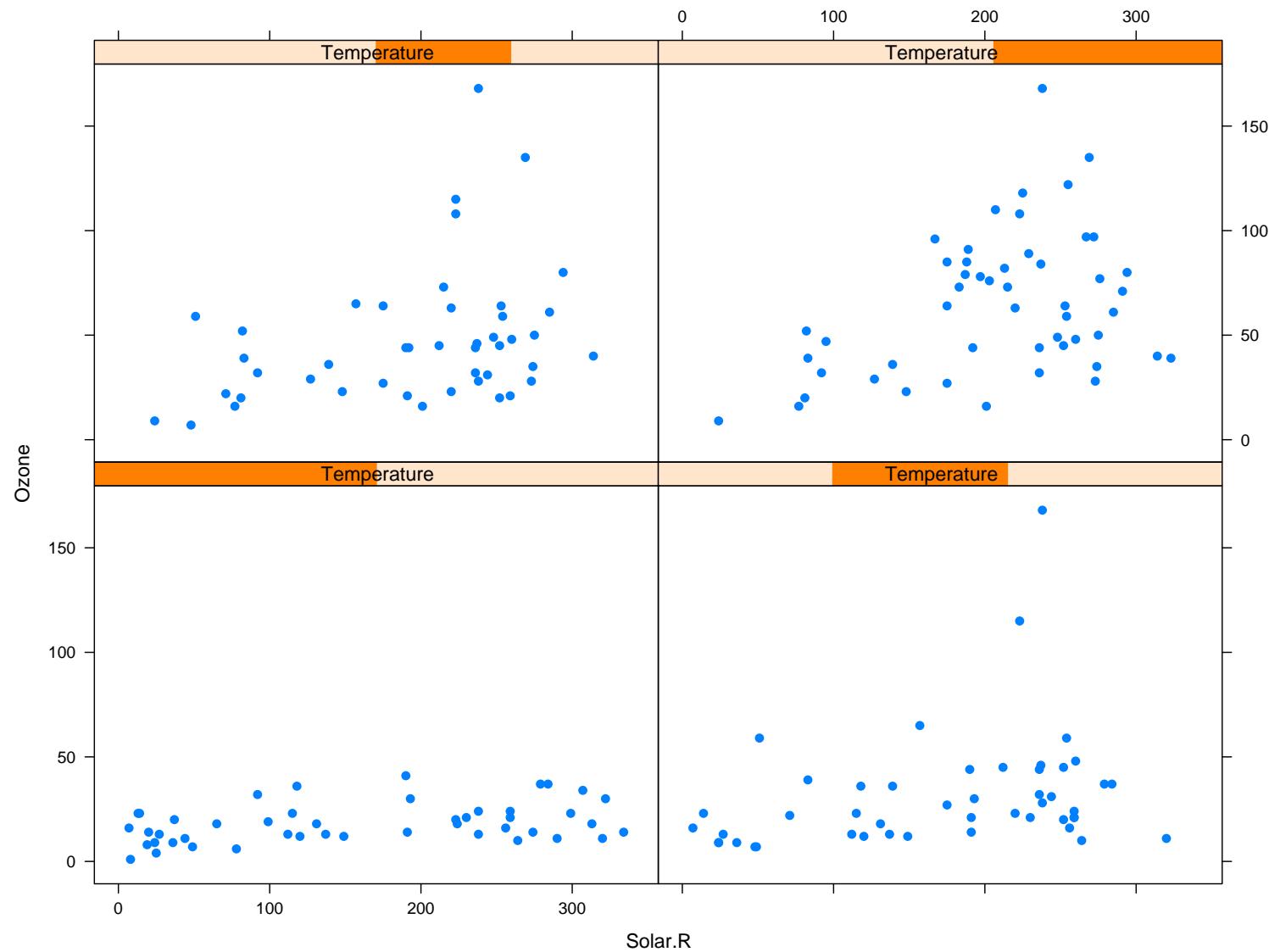


Table or Graph?

- Tables best suited for looking up specific information
- Graphs better for perceiving trends, making comparisons and predictions
- Ref. Gelman *et al.* (*Am Stat* 2002)

BIOS 662 Fall 2018

Random Number Generation

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Uses of random numbers in statistics
- Generation of random numbers from the uniform distribution
- Generation of random numbers from the normal distribution
- Random numbers from other distributions
- Controlling the sequence in SAS

Random Numbers in Statistics

Many statistical applications rely on a stream of random numbers:

- Simulations
- Sampling
- Randomization
- Re-sampling methods (e.g. bootstrap)
- Multiple imputation
- Markov Chain Monte Carlo methods

Generation of Uniform Random Numbers

- Properties required of a sequence of independent and identically distributed (iid) uniform (0, 1) random numbers:
 - uniformity
 - independence
 - identically distributed
- Mechanical methods
- <http://www.random.org/>

“RANDOM.ORG offers *true* random numbers to anyone on the Internet. The randomness comes from atmospheric noise, which for many purposes is better than the pseudo-random number algorithms typically used in computer programs.”

Generation of Uniform Random Numbers

- Statistical software packages provide one or more *pseudo*-random number generators
- A sequence of pseudo-random numbers is:
 - deterministic
 - reproducible
 - finite
- uniform random numbers usually uniform on $(0, 1)$
- if $X \sim U(0, 1)$,
let $Y = a + bX$,
then $Y \sim U(a, a + b)$

Generation of Uniform Random Numbers

- Multiplicative congruential generator

$$x_i = ax_{i-1} \pmod{m}, i = 1, 2, 3, \dots$$

- mod is the modulo (or modulus) operator, giving the remainder after dividing by m
- x_0 is called the *seed*
- the *cycle length* is the number of distinct values of x_i before the sequence starts to repeat itself
- the sequence is of full period if the cycle contains all the numbers in $\{1, 2, 3, \dots, (m - 1)\}$
- different values of x_0 correspond to starting at different points in the cycle

Uniform Random Number Generator in SAS

- CALL RANUNI(seed,x);
- x = RANUNI(seed);
- IML function: x = UNIFORM(seed);
 - x is the value returned by the function
 - seed is a non-negative integer $< 2^{31} - 1$
 - if seed = 0, time of day used as seed for the stream
 - in IML the seed can be a matrix
 - CALL RANUNI gives greater control of the seed and random number streams than RANUNI

Uniform Random Number Generator in SAS

- “ x is generated from the uniform distribution on the interval $(0,1)$, using a prime modulus multiplicative generator with modulus $2^{31} - 1$ and multiplier 397204094.”
- Cycle length is $2^{31} - 2 = 2.147 \times 10^9$
- Program on next page ran for 16 minutes, 42 seconds, on a PC, September 2012

Uniform Random Number Generator in SAS

```
proc iml;  
  n=1;  
  seed=83763;  
  x=ranuni(seed);  
  print x;  
  y=ranuni(seed);  
  do until (x=y);  
    n=n+1;  
    y=ranuni(seed);  
  end;  
  print y;  
  print n;  
quit;  
run;
```

n

2.14748E9

Generation of $N(0, 1)$ Random Numbers

- Box-Muller transformation: If u_1 and u_2 are independent random numbers from $U(0, 1)$, set:

$$z_1 = \sqrt{-2 \ln(u_1)} \cos(2\pi u_2)$$

$$z_2 = \sqrt{-2 \ln(u_1)} \sin(2\pi u_2)$$

then z_1 and z_2 are independent random numbers from $N(0, 1)$

- If z is a random number from $N(0, 1)$,
let $x = \mu + \sigma z$,
then x is a random number from $N(\mu, \sigma^2)$

$N(0, 1)$ Random Number Generator in SAS

- CALL RANNOR(seed, x);
- x = RANNOR(seed);
- IML function: x = NORMAL(seed);
 - x is the value returned by the function
 - seed is a non-negative integer $< 2^{31} - 1$
 - if seed = 0, time of day used as seed for the stream
 - in IML the seed can be a matrix
 - SAS uses the Box-Muller transformation of RANUNI uniform variates

Random Number Generation in R

- Random numbers from the uniform distribution

`runif(n, min=0, max=1)`

– n is the number of observations to generate from

$$U(\text{min}, \text{max})$$

– Default generator is “Mersenne-Twister”, the twisted generalized feedback shift register algorithm of Matsumoto and Nishimura (1998), with cycle length $2^{19937} - 1 = 10^{500}$

– `set.seed(k)`, where k is an integer; if the seed is not set, clock time is used to generate a seed

- Random numbers from the normal distribution

`rnorm(n, mean = 0, sd = 1)`

Another Random Number Generator in SAS

- SAS has a random number generator that uses the “Mersenne-Twister” with cycle length $2^{19937} - 1 = 10^{500}$.
- Syntax: `RAND('dist', parm-1, ..., parm-k)`
`dist` indicates the distribution from which to generate the random number; there are around 20 distributions available to be called by the function.
`parm-1, ..., parm-k` are parameters that need to be specified for the distribution; the number of parameters depends on the distribution.
- For the Normal: `RAND('NORMAL', mean, std_dev)`
- For the standard Normal: `RAND('NORMAL', 0, 1)`
or just `RAND('NORMAL')`

Random Numbers From Other Distributions

- Built-in functions for many generators in SAS, R, etc.
- Inverse transformation method
 - If the cumulative distribution function $F(x)$ can be written in closed form, set $u = F(x)$ and solve for x . Then, generate u from $U(0, 1)$ and calculate x
- Example: exponential distribution

$$f(x) = \lambda e^{-\lambda x} \quad \text{and} \quad F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

Set $u = F(x) = 1 - e^{-\lambda x}$.

Solving for x : $x = -\frac{1}{\lambda} \ln(1 - u)$.

Because $u \sim U(0, 1) \Rightarrow 1 - u \sim U(0, 1)$

we can use $x = -\frac{1}{\lambda} \ln(u)$

Setting the Seed in SAS

- It is usually desirable to use a specified seed rather than `seed = 0` because
 - debugging a program is easier
 - results are repeatable (e.g. for an audit)
- SAS makes this difficult when random numbers are needed in multiple data steps, such as a simulation repeating a data step multiple times

```
%let seed0=97231;  
%let sampsize=4;  
%let nreps=2;  
  
data begindat;  
  do i=1 to &sampsize;  
    output;  
  end;
```

```
%macro simulation1(reps=);
  %do i=1 %to &reps;
    data sim1; set begindat;
    x=ranuni(&seed0);
    proc print data=sim1;
  %end;
%mend simulation1;

%simulation1(reps=&nreps);
```

Obs i x

1	1	0.09563
2	2	0.69591
3	3	0.31711
4	4	0.58969

Obs i x

1	1	0.09563
2	2	0.69591
3	3	0.31711
4	4	0.58969

```
%macro simulation1(reps=);
  %do i=1 %to &reps;
    data sim1; set begindat;
    x=ranuni(&seed0+&i);
    proc print data=sim1;
  %end;
%mend simulation1;

%simulation1(reps=&nreps);
```

Obs	i	x
1	1	0.28059
2	2	0.66600
3	3	0.71693
4	4	0.84909

Obs	i	x
1	1	0.46555
2	2	0.63609
3	3	0.11675
4	4	0.10849

- Although the example on the previous page works, it is unsatisfactory because it is not clear how the sequence jumps around in the cycle
- How about putting the seed in the call statement?

```
call ranuni(&seed0,x);
```

```
ERROR 135-185: Attempt to change the value of the constant 97231  
in the RANUNI subroutine call.
```

- Try assigning the macro variable `&seed0` to another variable first

```
%macro simulation2(reps=);
  %do i=1 %to &reps;
    data sim2; set begindat;
    seed=&seed0;
    call ranuni(seed,x);
    proc print data=sim2;
  %end;
%mend simulation2;
```

```
%simulation2(reps=&nreps);
```

Obs	i	seed	x
1	1	205356066	0.095626
2	2	205356066	0.095626
3	3	205356066	0.095626
4	4	205356066	0.095626

Obs	i	seed	x
1	1	205356066	0.095626
2	2	205356066	0.095626
3	3	205356066	0.095626
4	4	205356066	0.095626

```
%macro simulation3(reps=);
  %do i=1 %to &reps;

data sim3;
  set begindat end=eof;

  retain seed &seed0;

  call ranuni(seed,x);

  if eof then do;
    call symput('seed0',put(seed,best.));
  end;

proc print data=sim3;

%end;
%mend simulation3;

%simulation3(reps=&nreps);
```

Obs	i	seed	x
1	1	205356066	0.09563
2	2	1494458509	0.69591
3	3	680978604	0.31711
4	4	1266346632	0.58969

Obs	i	seed	x
1	1	730678039	0.34025
2	2	1730858377	0.80599
3	3	876140027	0.40798
4	4	1971178368	0.91790

```
%let sampsize=8;
```

Obs	i	seed	x
1	1	205356066	0.09563
2	2	1494458509	0.69591
3	3	680978604	0.31711
4	4	1266346632	0.58969
5	5	730678039	0.34025
6	6	1730858377	0.80599
7	7	876140027	0.40798
8	8	1971178368	0.91790

BIOS 662 Fall 2018

Statistical Inference: Populations and Samples

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Random Variables

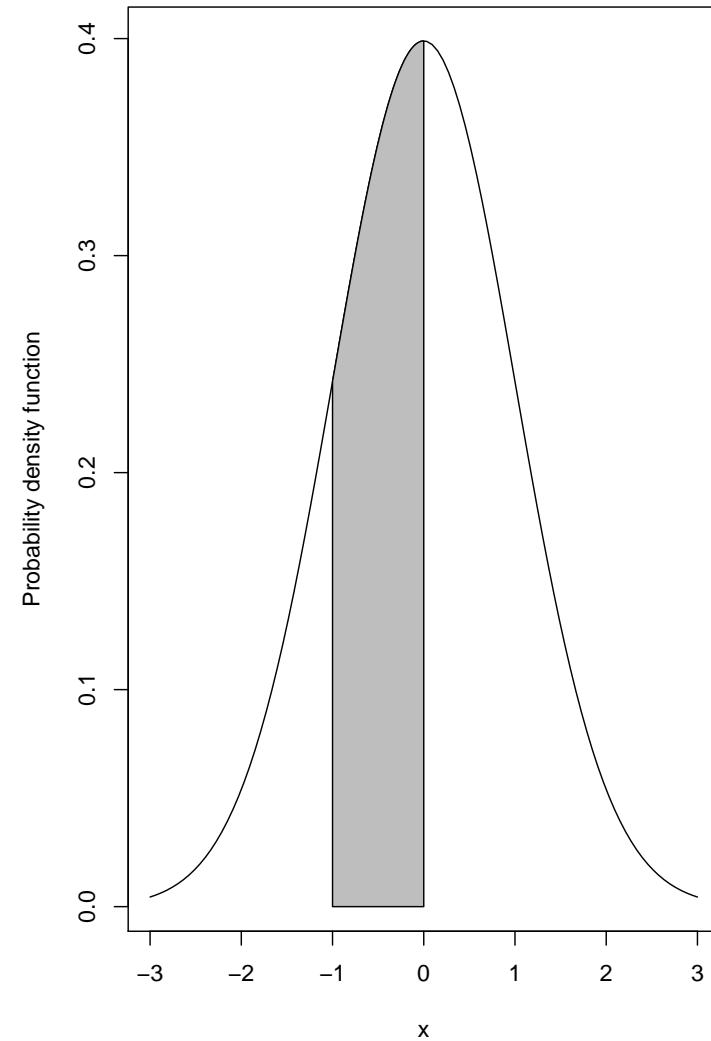
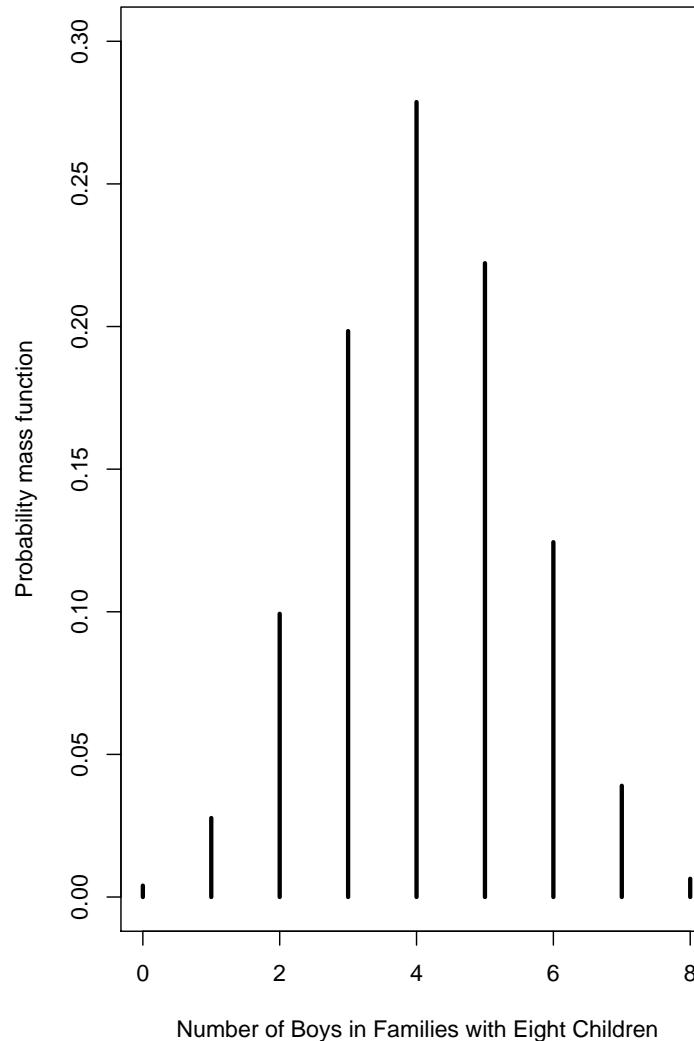
- *Random sample*: result of independently selecting elements at random from a population
- Definition 4.8. A *random variable* is a variable associated with a random sample

P.V. Rao (1998, p 786): A *random variable* is a variable whose value is determined by the observed characteristics of an item randomly selected from a population

Probability Functions

- Definition 4.9. The *probability mass function* (pmf) is a function that for each possible value of a discrete random variable takes on the probability of that value occurring
- Definition 4.10. The *probability density function* (pdf) is a curve that specifies, by means of the area under the curve over an interval, the probability that a continuous random variable falls within the interval

Probability Functions



Cumulative Distribution Function

- Definition 4.9. The *cumulative distribution function* for a random variable X is

$$F(x) = \Pr[X \leq x]$$

- If X is discrete,

$$F(x) = \sum_{y \leq x} p_X(y)$$

where p_X is the pmf of X

- If X is continuous,

$$F(x) = \int_{-\infty}^x f(y) dy$$

where f is the pdf of X

Population Quantile

- Intuitive definition:

The p^{th} quantile of X , say ζ_p , should be such that

$$F(\zeta_p) = \Pr[X \leq \zeta_p] = p$$

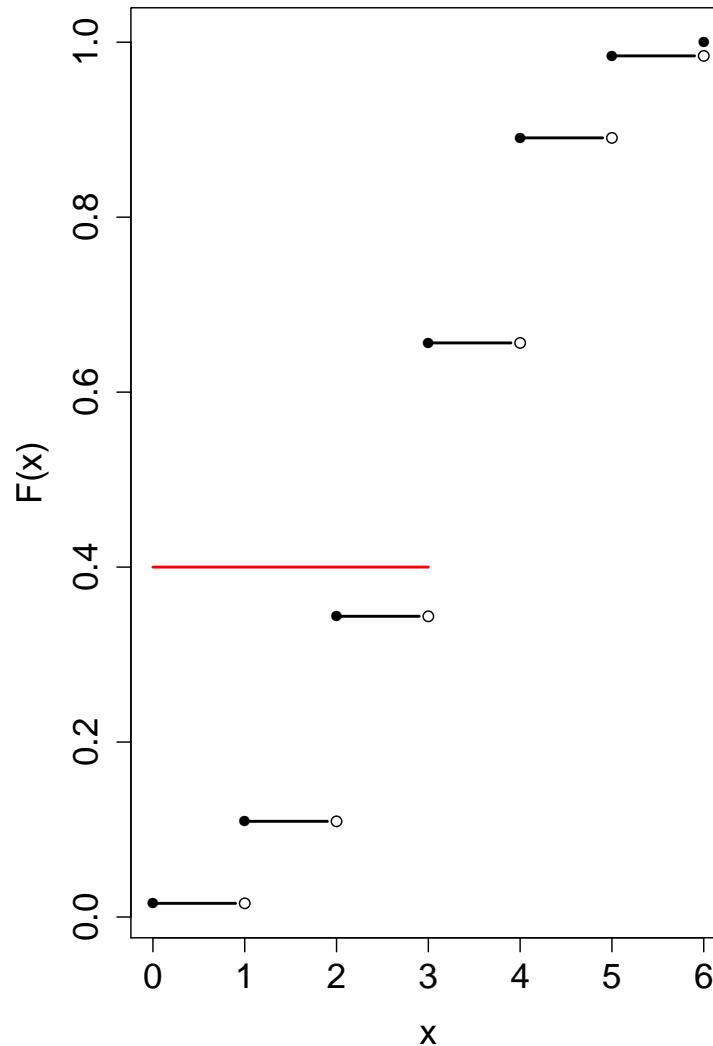
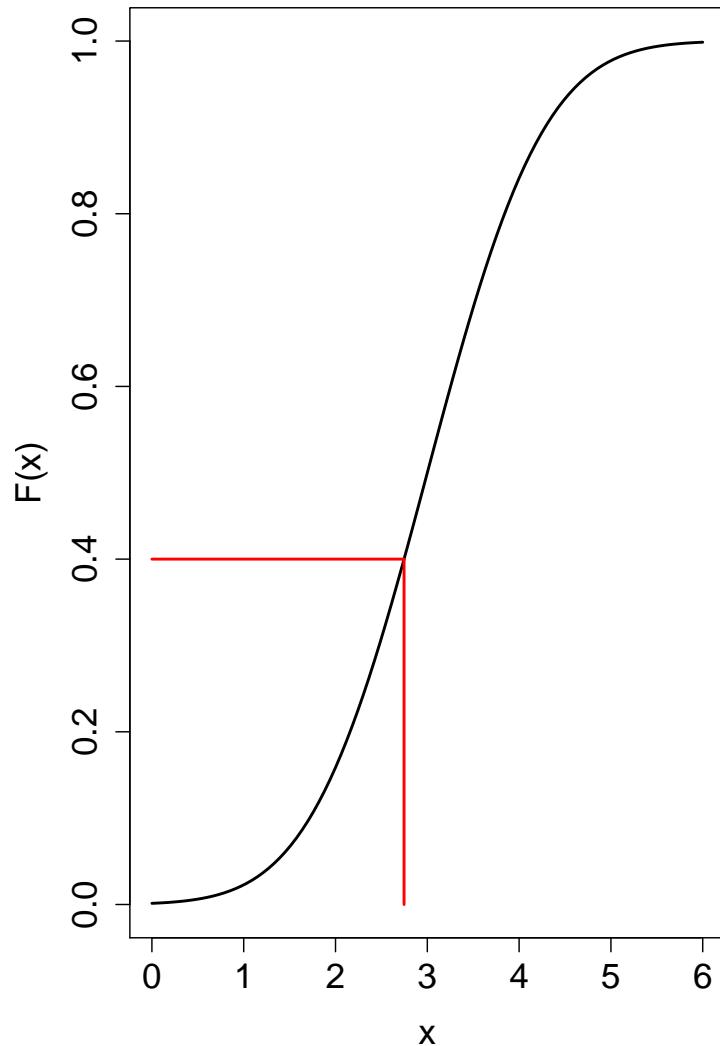
- Formally:

$$\zeta_p = \inf\{x : F(x) \geq p\}$$

- If F is continuous

$$F(\zeta_p) = p$$

Quantiles: Example



Mean and Variance

- Mean or expected value of X

- If X is discrete,

$$\mu = E(X) = \sum_x x p_X(x)$$

- If X is continuous,

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

- Variance

$$\sigma^2 = \text{Var}(X) = E\{(X - \mu)^2\}$$

- $E(g(X)) = \sum_x g(x) p_X(x)$ (discrete)
 $= \int_{-\infty}^{\infty} g(x) f(x) dx$ (continuous)

Skewness and Kurtosis

- Skewness

$$\alpha_3 = \frac{E\{(X - \mu)^3\}}{\sigma^3}$$

- Kurtosis

$$\alpha_4 = \frac{E\{(X - \mu)^4\}}{\sigma^4}$$

Parameters and Statistics

- Definition: A *parameter* is a numerical characteristic of a population
- Definition: A *statistic* is a numerical characteristic of a sample
- Notation: Greek letters typically denote parameters;
Latin / English letters denote statistics
- Example:
 - μ population mean; σ^2 population variance
 - \bar{Y} sample mean; s^2 sample variance

Parameters and Statistics

- Parameters are fixed constants
- Statistics are random variables
- Statistics have probability distributions
- We will use statistics and probability theory to draw conclusions (inference) about parameters

Sampling Distributions

- Definition 4.15. The probability function of a statistic is called the *sampling distribution of the statistic*
- For example, when sampling from a population, the sample mean \bar{Y} is a random variable because its value depends on chance, namely, on which sample is obtained
- The probability distribution of the random variable \bar{Y} is called the *sampling distribution of the mean*

Sampling Distributions

- Result 4.1. If a random variable Y has a population mean μ and a population variance σ^2 , the sampling distribution of the mean (\bar{Y}) has mean μ and variance σ^2/n
- Definition 4.16. The standard deviation of the sampling distribution is called the *standard error*
- For example, the standard error of \bar{Y} is σ/\sqrt{n}

Normal or Gaussian Distribution

- PDF:

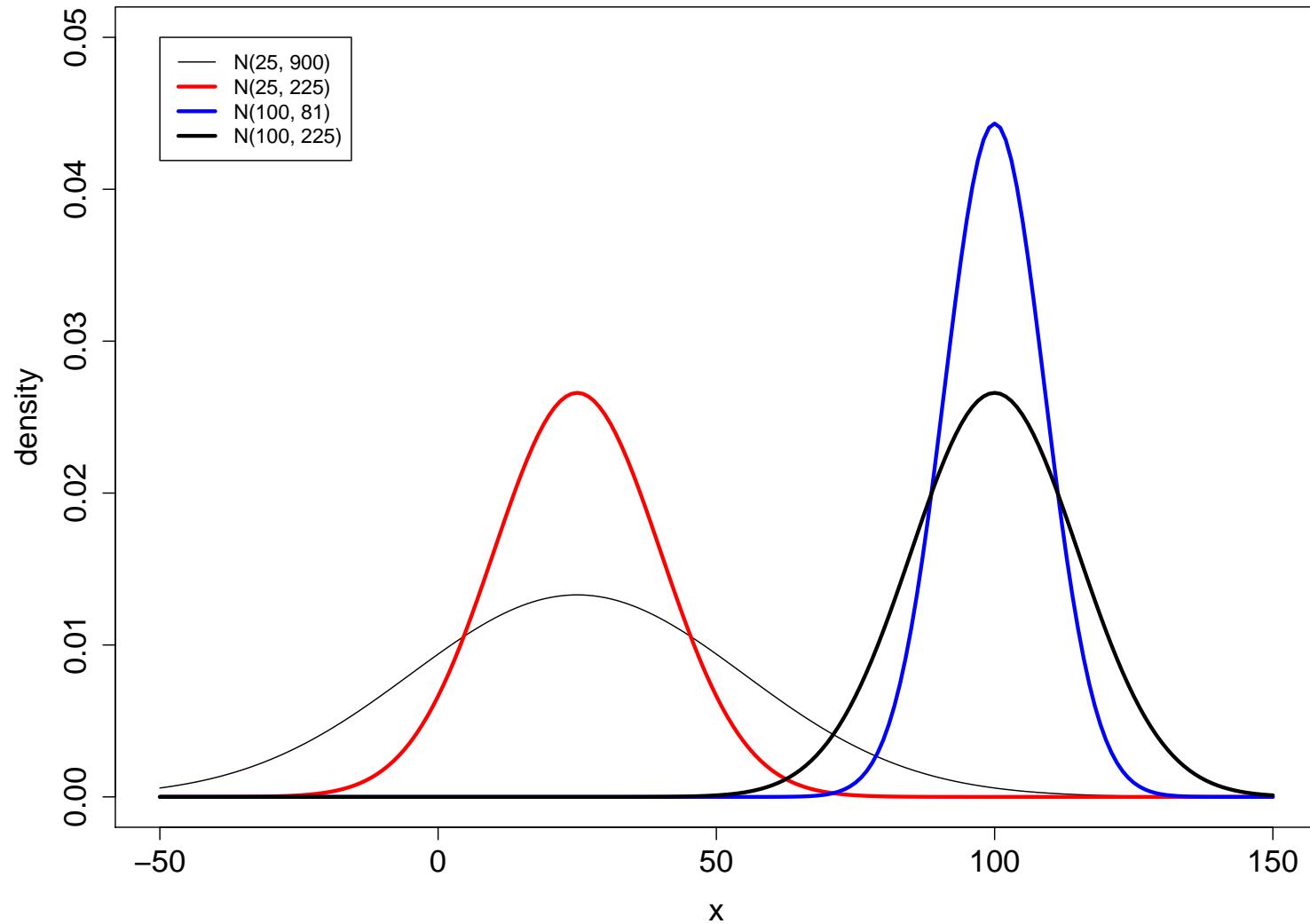
$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

- CDF:

$$F(x; \mu, \sigma) = \int_{-\infty}^x f(y; \mu, \sigma) dy$$

- μ mean, σ^2 variance
- $X \sim N(\mu, \sigma^2)$ [beware $X \sim N(\mu, \sigma)$]

Normal Distribution



Standard Normal Distribution

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

- PDF:

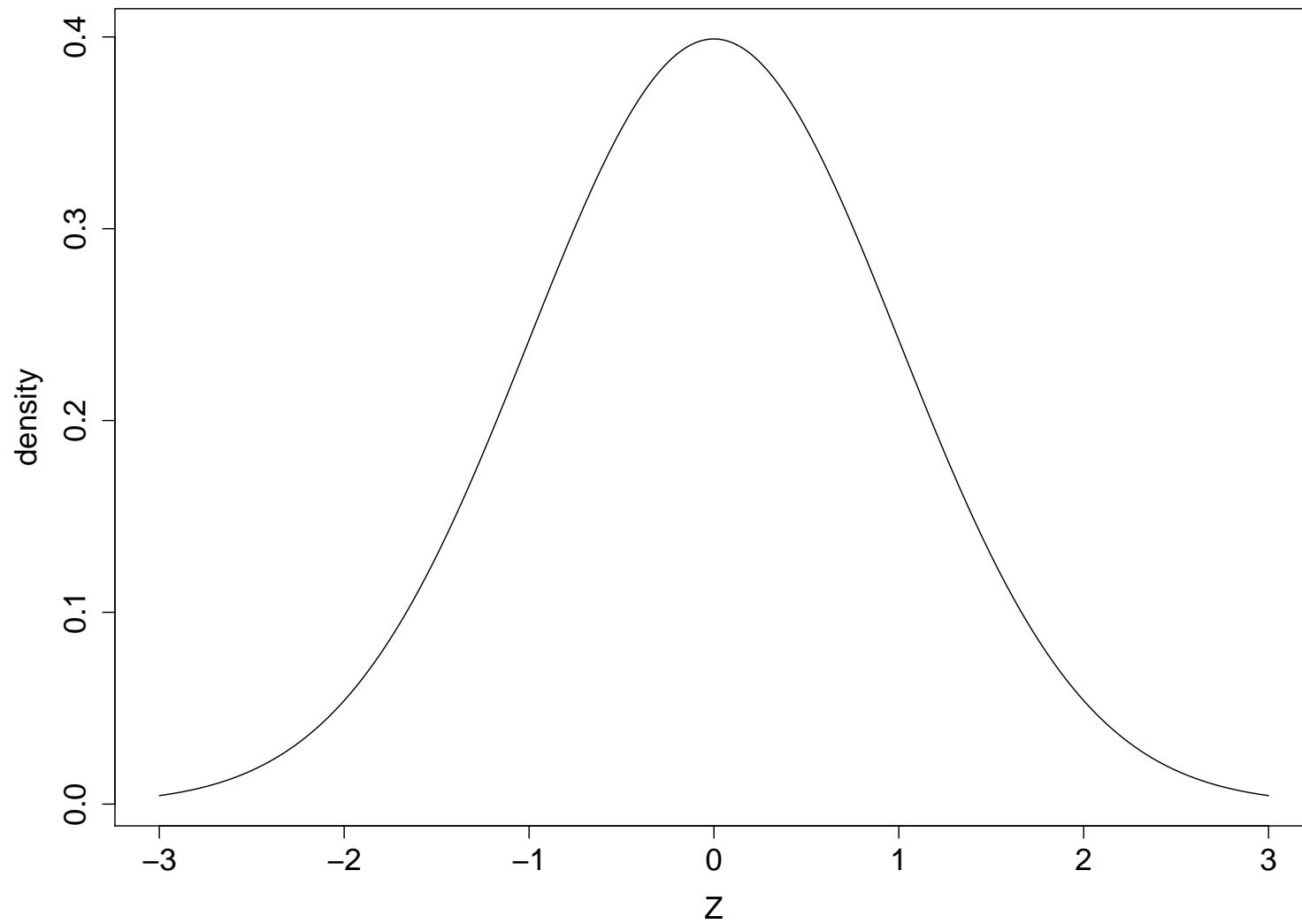
$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

- CDF:

$$\Phi(z) = \int_{-\infty}^z \phi(y) dy$$

- $N(0, 1)$ is the *standard normal distribution*

Standard Normal Distribution



Properties of the Standard Normal Distribution

- A random variable with pdf f is *symmetric* about μ if

$$f(\mu + x) = f(\mu - x) \text{ for all } x$$

- $Z \sim N(0, 1)$ is symmetric about 0

$$\phi(z) = \phi(-z) \text{ for all } -\infty < z < \infty$$

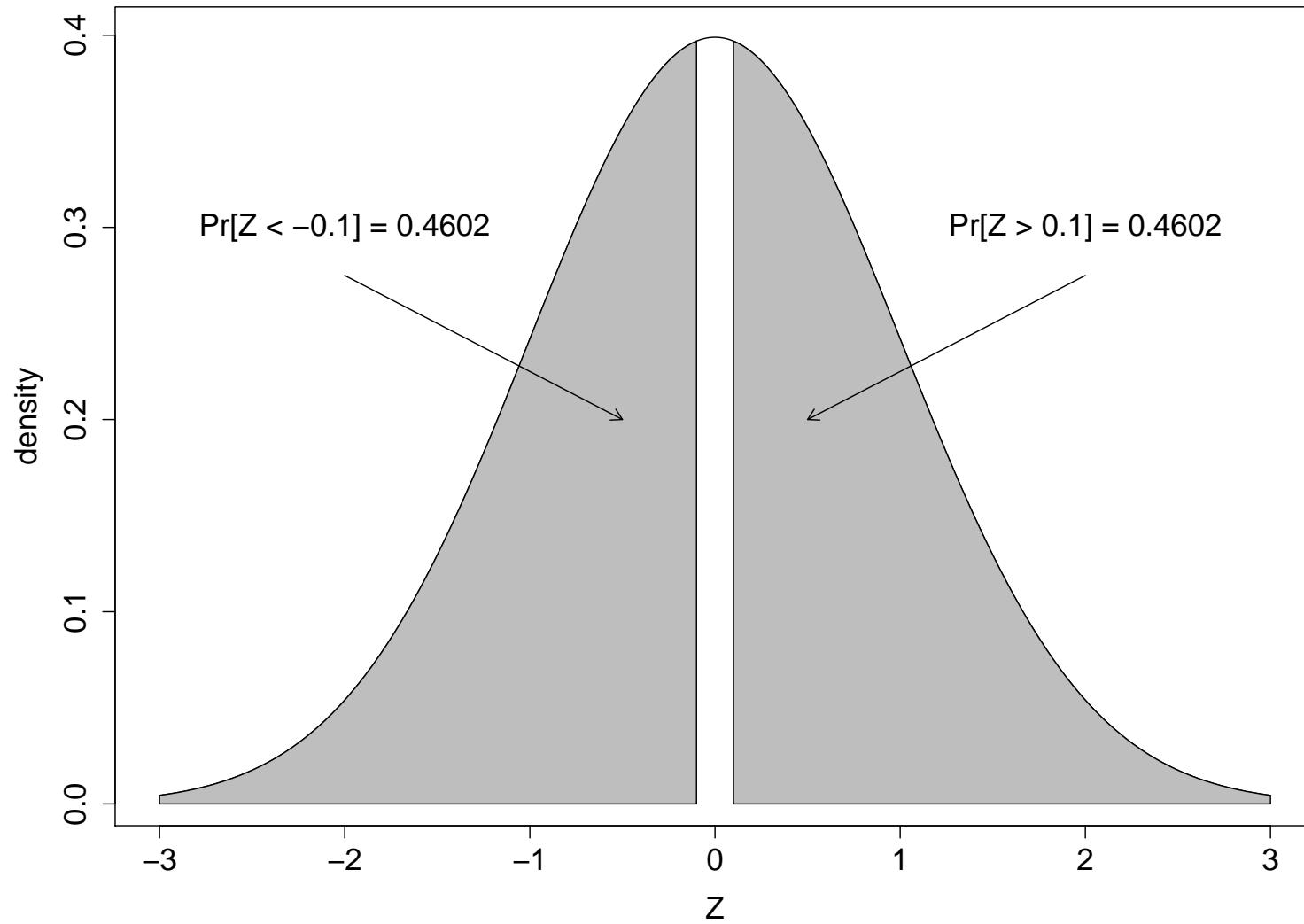
- Thus

$$\Pr[Z \leq -z] = \Pr[Z \geq z]$$

i.e.

$$\Phi(-z) = 1 - \Phi(z)$$

Standard Normal Distribution



Standard Normal Distribution

- R

```
> pnorm(-0.1,0,1)
[1] 0.4601722
> 1-pnorm(0.1,0,1)
[1] 0.4601722
> qnorm(0.4601722,0,1)
[1] -0.0999999
```

Standard Normal Distribution

- SAS

```
data normal;  
x=probnorm(-0.1);  
y=cdf('NORMAL',-0.1,0,1);  
z=quantile('NORMAL',0.4601722);
```

```
proc print data=normal;
```

Obs	x	y	z
1	0.46017	0.46017	-0.100000

Properties of a Random Variable

- Let X be a random variable
- Suppose

$$Y = aX + b$$

where a and b are constants

- Then

$$E(Y) = aE(X) + b$$

$$\text{Var}(Y) = a^2 \text{Var}(X)$$

- If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then

$$Y \sim N(a\mu + b, (a\sigma)^2) = N(a\mu + b, a^2\sigma^2)$$

Conversion to Standard Normal

- Suppose $Y \sim N(\mu, \sigma^2)$

- Let

$$Z = \frac{Y - \mu}{\sigma}$$

- Then

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

- In words: any normally distributed random variable can be standardized by subtracting its mean and dividing by its standard deviation

Computation of Probabilities

- Suppose $Y \sim N(\mu, \sigma^2)$

- Let

$$Z = \frac{Y - \mu}{\sigma}$$

- Then

$$\Pr[a < Y < b] = \Pr\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right]$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Table 1 (text, p 818): Standard Normal Distribution

Let Z be a normal random variable with mean zero and variance one. For selected values of z , three values are tabled: (1) the two-sided p -value, or $\Pr[|Z| \geq z]$; (2) the one-sided p -value, or $\Pr[Z \geq z]$; and (3) the cumulative distribution function at z , or $\Pr[Z \leq z]$.

z	Two-sided	One-sided	Cum-dist.
0.00	1.0000	.5000	.5000
0.05	.9601	.4801	.5199
0.10	.9203	.4602	.5398
0.15	.8808	.4404	.5596
0.20	.8415	.4207	.5793
0.25	.8026	.4013	.5987
0.30	.7642	.3821	.6179
0.35	.7263	.3632	.6368
0.40	.6892	.3446	.6554
0.45	.6527	.3264	.6736
⋮			
1.00	.3173	.1587	.8413
1.33	.1835	.0918	.9082
1.64	.1010	.0505	.9495
1.96	.0500	.0250	.9750
2.00	.0455	.0288	.9772
2.58	.0099	.0049	.9951

Example

- Intraocular pressure (IP) is used to diagnose glaucoma
- Assume IP is normally distributed with mean $\mu = 16 \text{ mmHg}$ and variance $\sigma^2 = 9 \text{ mmHg}^2$
- If pressure greater than 20 mmHg is considered abnormal, what proportion of the population is abnormal?

$$\begin{aligned}\Pr[X > 20] &= \Pr\left[\frac{X-16}{3} > \frac{20-16}{3}\right] \\ &= \Pr[Z > 1.33] = 1 - \Phi(1.33) \\ &= 1 - 0.9082 = 0.0918\end{aligned}$$

Example (continued)

- What proportion of the population has IP between 4 and 18?

$$\begin{aligned}\Pr[4 < X < 18] &= \Pr\left[\frac{4-16}{3} < \frac{X-16}{3} < \frac{18-16}{3}\right] \\ &= \Pr[-4 < Z < 2/3] \\ &= \Phi(2/3) - \Phi(-4) \\ &= \Phi(2/3) - 1 + \Phi(4) \\ &= 0.7475 - 1 + 0.99997 = 0.7475\end{aligned}$$

Assessing Normality

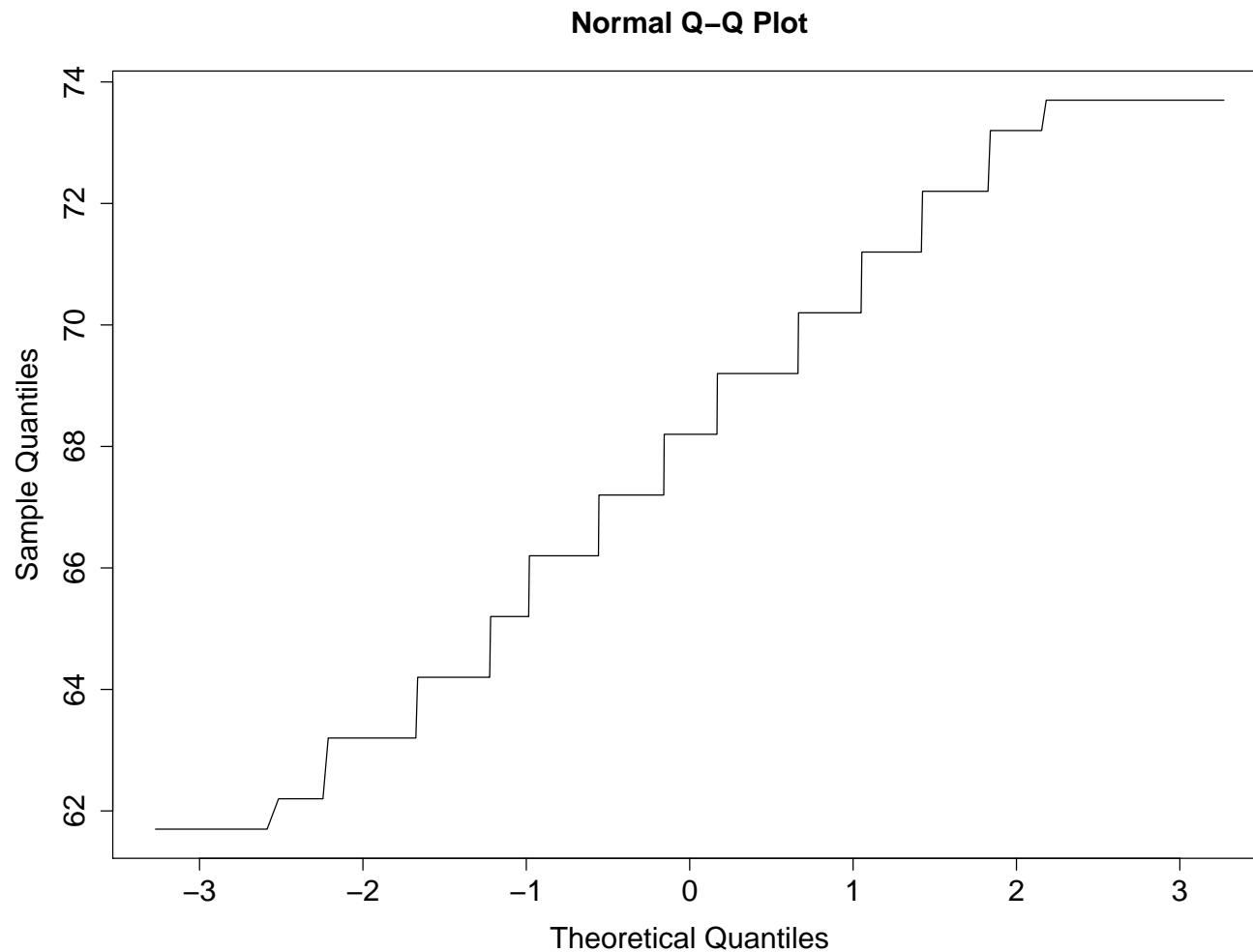
- How do we assess whether the normal distribution model is a reasonable fit for a particular set of data?
- One graphical approach: quantile-quantile (QQ) plot
- Plot quantiles of the observed data distribution versus the quantiles of the normal distribution
- Straight line indicates normality assumption reasonable

QQ Plot Example

- Table 4.3 from text (p. 81)

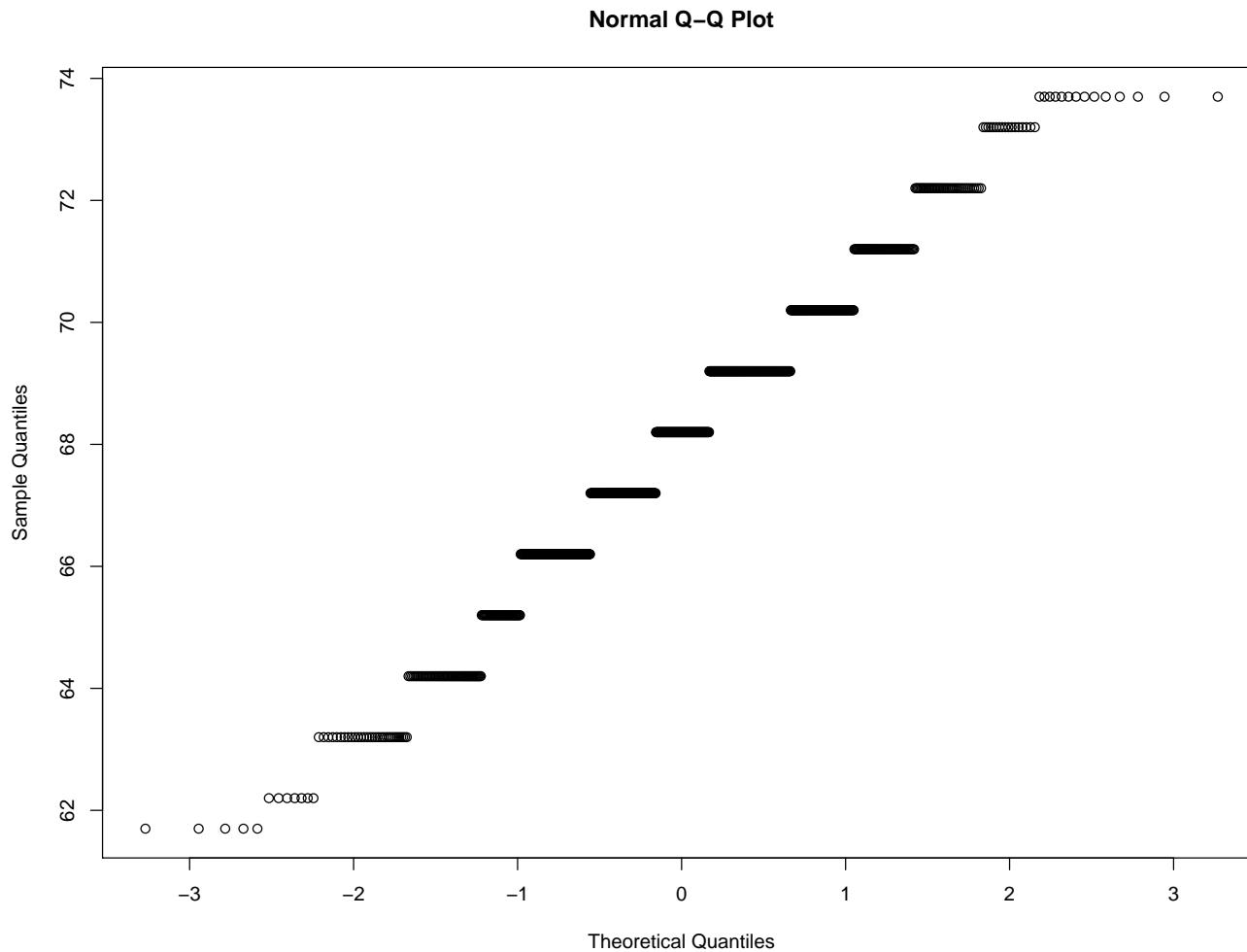
Endpoint	Frequency	Cumulative Percentage
61.7	5	0.5
62.2	7	1.3
63.2	32	4.7
64.2	59	11.1
65.2	48	16.3
66.2	117	28.9
67.2	138	43.8
68.2	120	56.7
69.2	167	74.7
70.2	99	85.3
71.2	64	92.2
72.2	41	96.7
73.2	17	98.5
73.7	14	100.0

R QQ Plot Example: Table 4.3 from Text



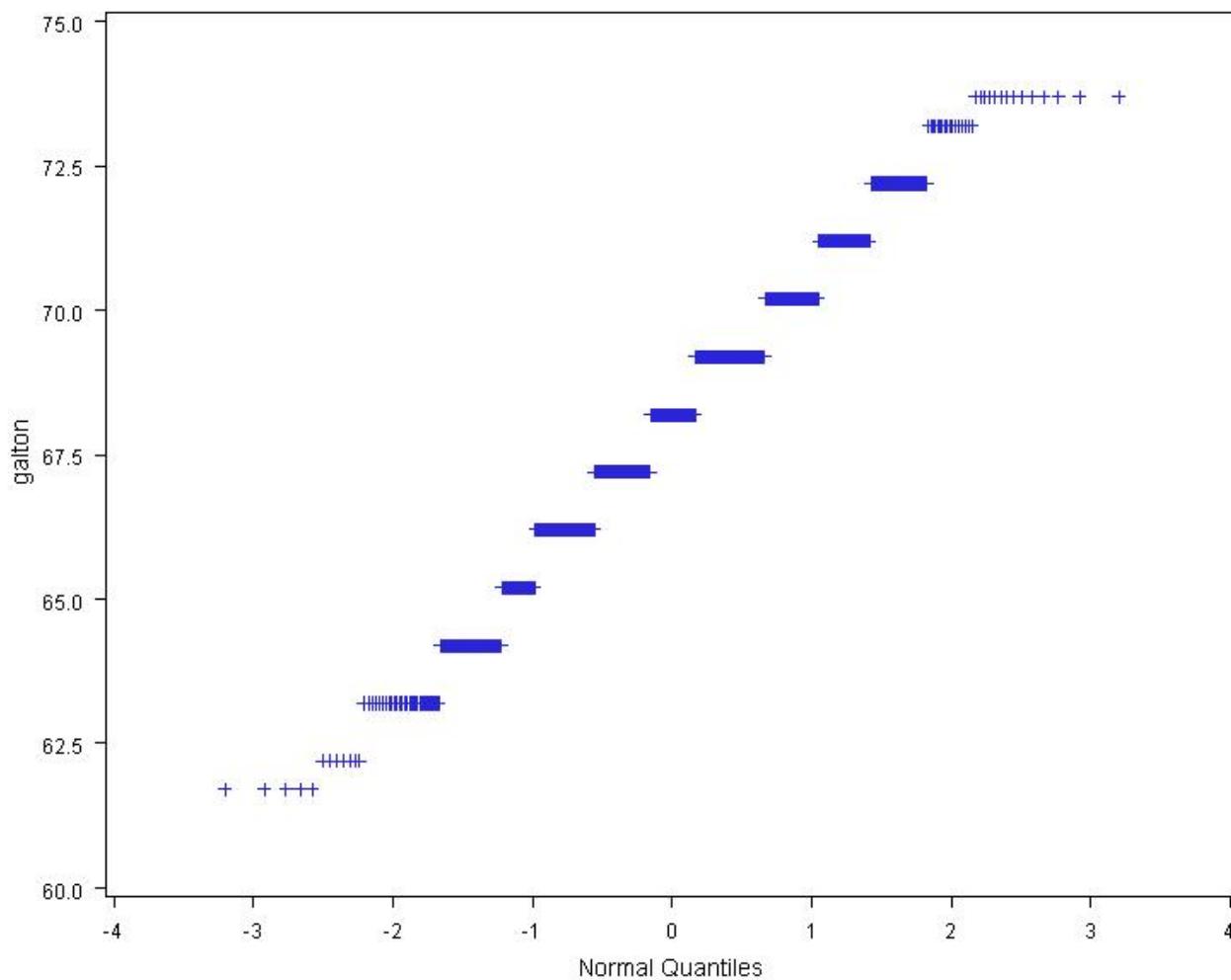
```
> qqnorm(galton,type="l")      # type="l" draws lines
```

R QQ Plot Example: Table 4.3 from Text



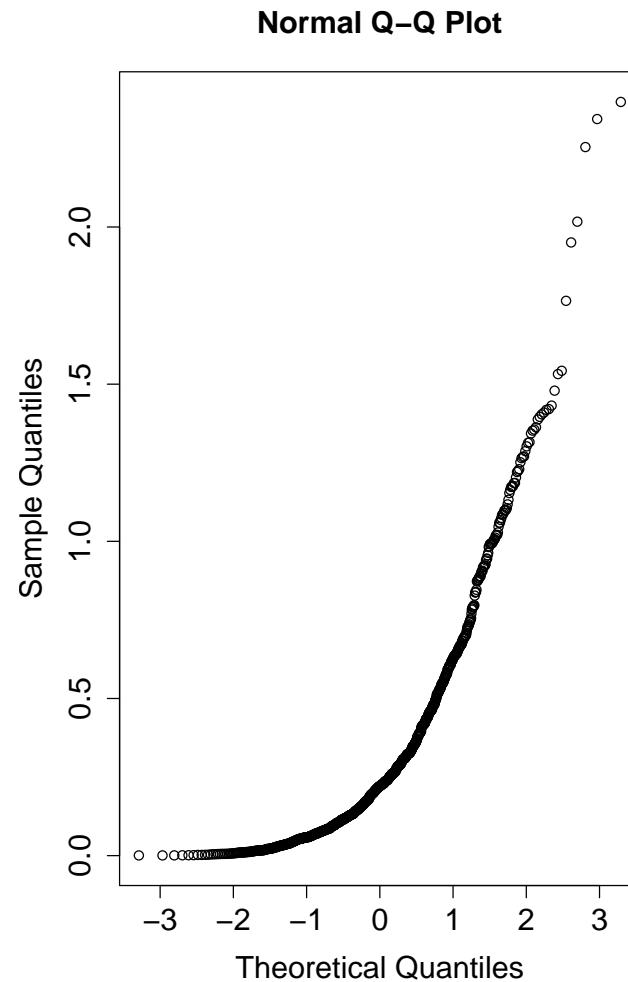
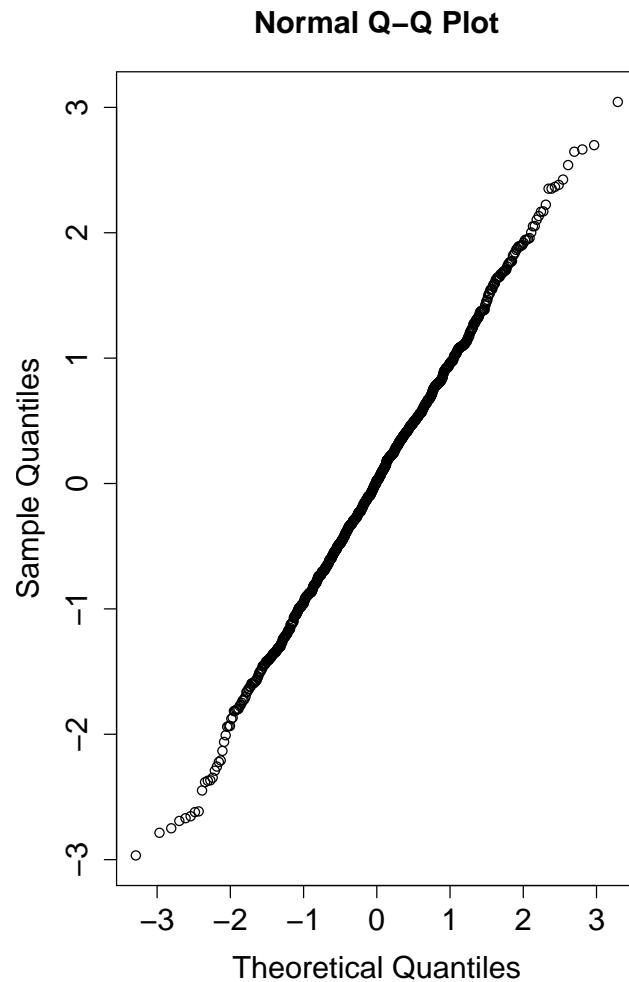
> qqnorm(galton)

SAS QQ Plot Example: Table 4.3 from Text



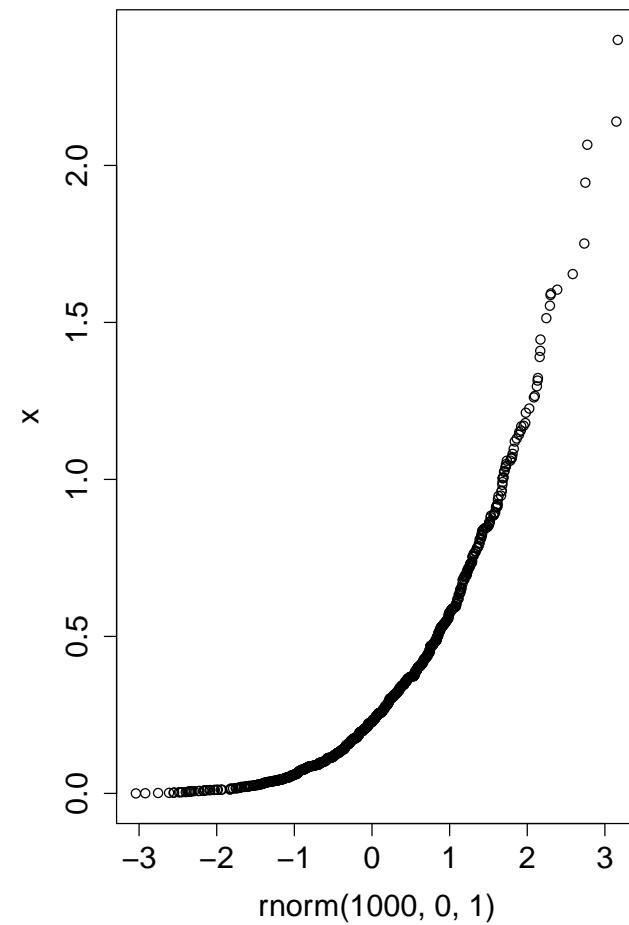
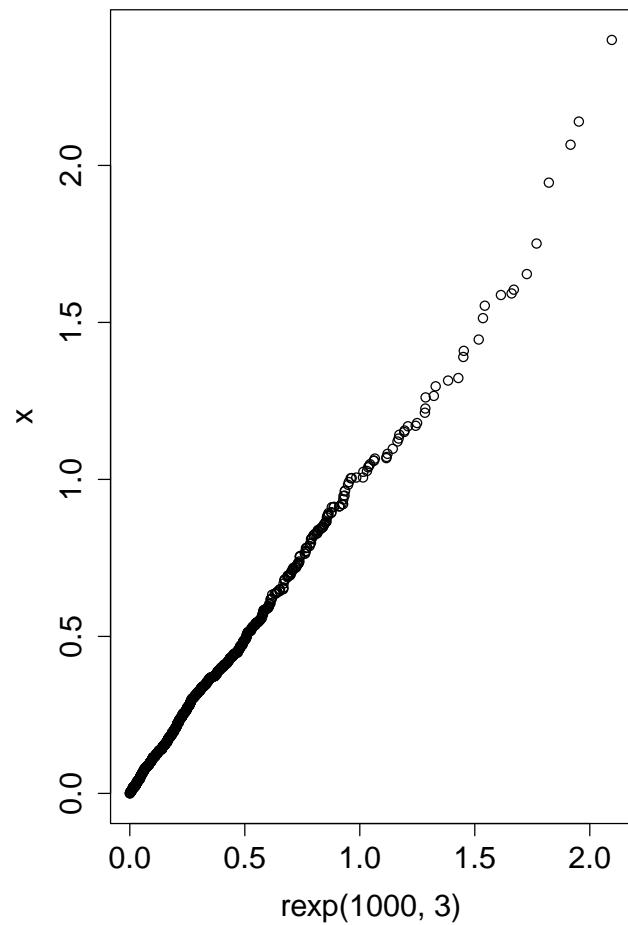
```
proc univariate;  
qqplot galton;
```

R QQ Plots



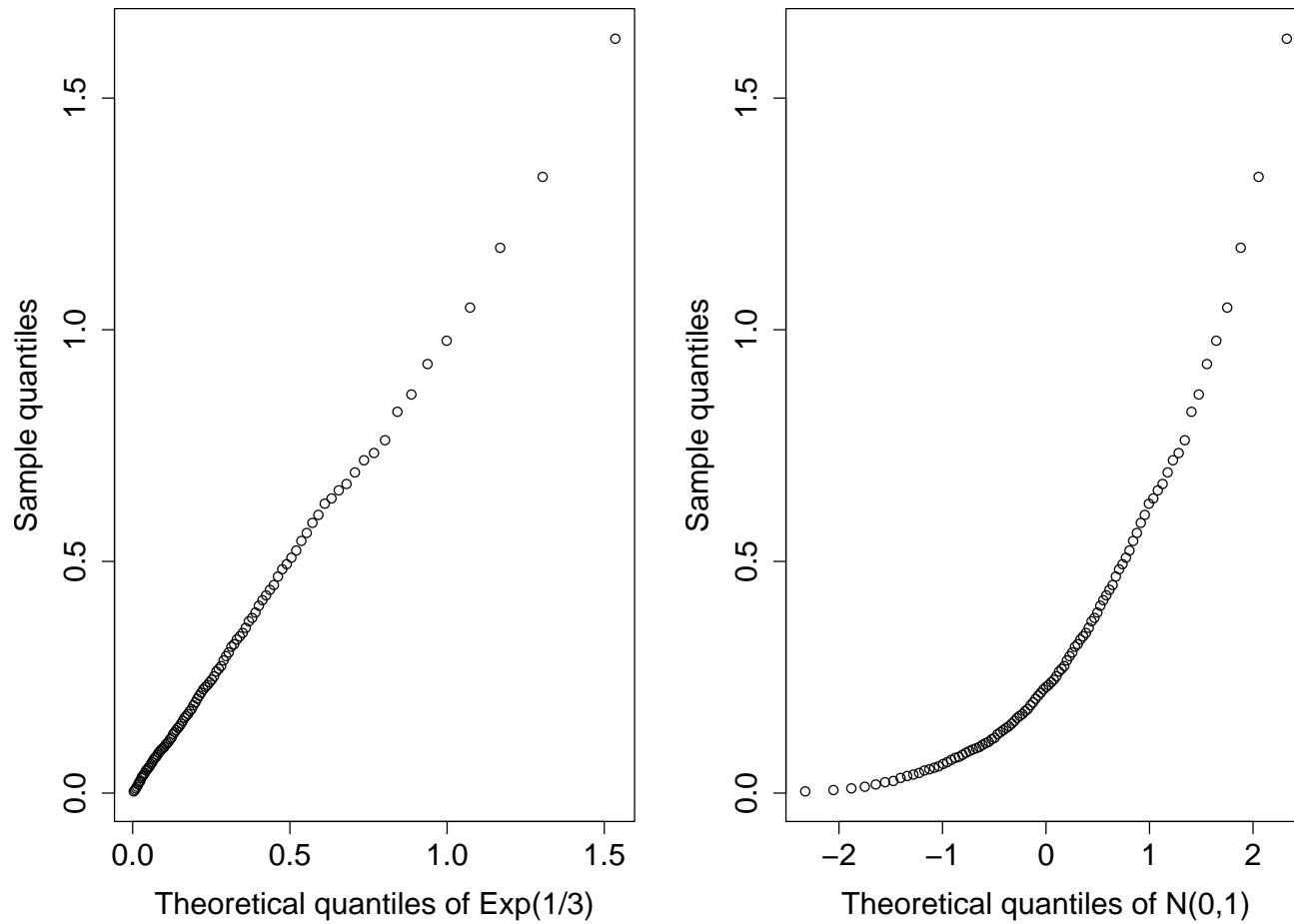
```
> par(mfcol=c(1,2)); qqnorm(rnorm(1000,0,1)); qqnorm(rexp(1000,3))
```

R QQ Plots



```
> x <- rexp(1000,3)
> par(mfcol=c(1,2)); qqplot(rexp(1000,3),x); qqplot(rnorm(1000,0,1),x)
```

R QQ Plots



```
> x <- rexp(1000,3); probs <- seq(0.01,0.99,length=99); par(mfcol=c(1,2))
> qx <- quantile(x,probs); tqexp <- qexp(probs,3); tqnorm <- qnorm(probs,0,1)
> plot(tqexp,qx,xlab="Theoretical quantiles of Exp(1/3)", ylab="Sample quantiles")
> plot(tqnorm,qx,xlab="Theoretical quantiles of N(0,1)",ylab="Sample quantiles")
```

Some Approximations for the Normal

- The interval $\bar{x} \pm s$ will contain approx 68% of the observations
- The interval $\bar{x} \pm 2s$ will contain approx 95% of the observations
- Assuming $Y \sim N(\mu, \sigma^2)$
$$\Pr[\mu - \sigma < Y < \mu + \sigma] = \Pr[-1 < Z < 1] = 0.6827$$
$$\Pr[\mu - 2\sigma < Y < \mu + 2\sigma] = \Pr[-2 < Z < 2] = 0.9545$$

Some Approximations

- Do these approximations hold for non-normal data?
Not in general.
- Consider $X \sim \text{Exp}(1/\lambda)$ such that $E(X) = \lambda$ and $\text{Var}(X) = \lambda^2$.

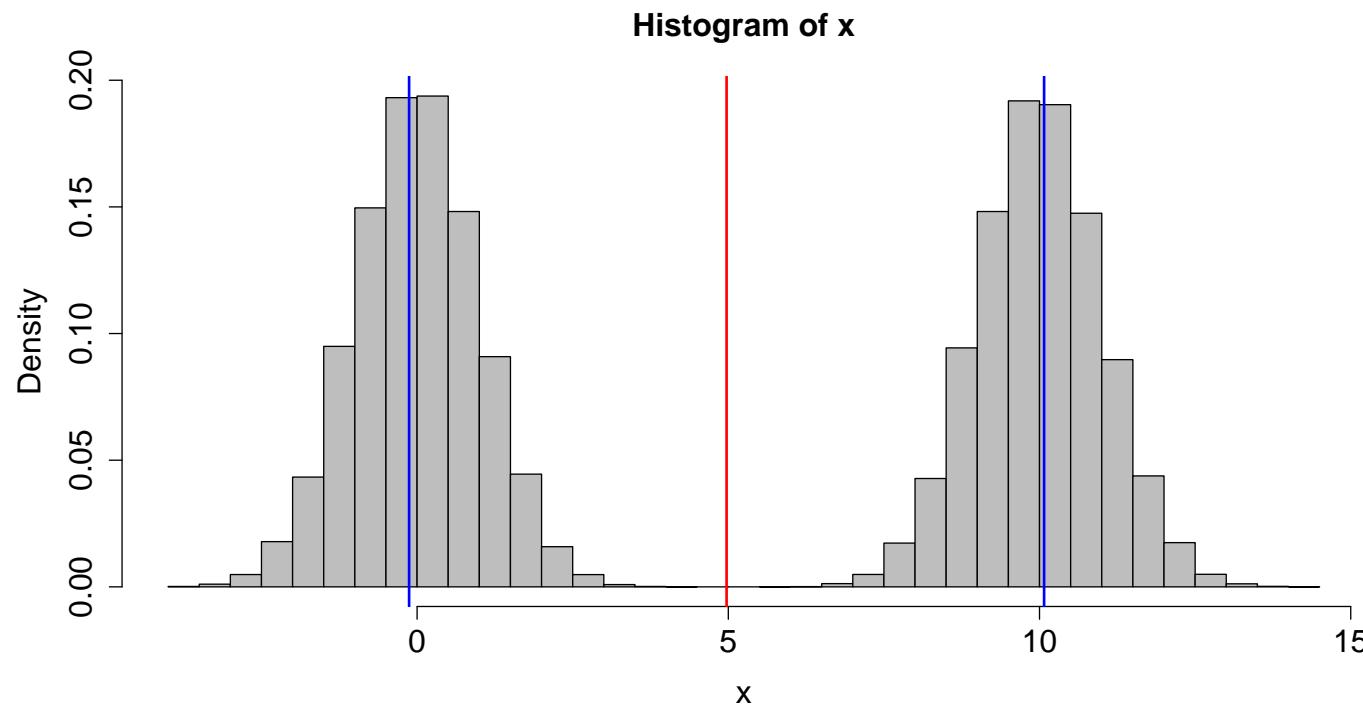
For $\lambda = 1/3$,

$$\Pr[0 \leq X \leq 2/3] = 0.86$$

Some Approximations

- Consider $X = WY + (1 - W)Z$
where $W \sim \text{Bernoulli}(1/2)$, $Y \sim N(10, 1)$,
and $Z \sim N(0, 1)$.

Can show $\Pr[\mu_X - \sigma_X \leq X \leq \mu_X + \sigma_X] \approx 0.54$



Some Approximations

- The following holds for any random variable Y with mean μ and variance σ^2

$$\Pr \left[\left| \frac{Y - \mu}{\sigma} \right| < K \right] = \Pr [\mu - K\sigma < Y < \mu + K\sigma] \geq 1 - \frac{1}{K^2}$$

$\forall K \geq 1$. This is *Chebyshev's inequality* (note typos in text on page 100)

- For example, if $K = 2$,

$$\Pr [\mu - 2\sigma < Y < \mu + 2\sigma] \geq 0.75$$

i.e. we would expect at least 75% of observations to be within two standard deviations of the mean *for any underlying distribution*

Central Limit Theorem (CLT)

- Let Y_1, Y_2, \dots, Y_n be independent and identically distributed (iid) random variables with

$$E(Y_i) = \mu \quad (\text{finite})$$

and

$$\text{Var}(Y_i) = \sigma^2 > 0$$

- Define

$$Z_n = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

- Then the distribution function of Z_n converges to the standard normal distribution function as $n \rightarrow \infty$.

Central Limit Theorem (CLT)

- In words - see Result 4.3 on page 84 of the textbook
- If a random variable Y has population mean μ and population variance σ^2 , then the sample mean \bar{Y} , based on n observations, is approximately normally distributed with mean μ and variance σ^2/n for sufficiently large n

Notes on the CLT

- The CLT applies to any distribution of the Y s
- The approximation improves as n gets large
- Check out the Rice Virtual Lab in Statistics

<http://onlinestatbook.com/rvls.html>

Result 4.2

- If Y is *normally distributed* with mean μ and variance σ^2 , then \bar{Y} , based on a random sample of n observations, is normally distributed with mean μ and variance σ^2/n .
- This is true regardless of sample size.

BIOS 662 Fall 2018

Point and Interval Estimation

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Introduction
- Confidence intervals (CIs) for the mean
 - Parametric, large sample
 - Bootstrap
- CI for quantiles
 - Exact
 - Large sample
- CI for variance

Inference

- *Inference*: Using statistics and probability theory to draw conclusions about parameters
- Two modes of inference:
 - **Estimation**: attempt to estimate value of parameter(s) and quantify uncertainty about these estimate(s)
 - **Hypothesis testing**: posit certain values for parameters and test whether the observed data are consistent with the hypothesis

Estimation

- *Estimand*: parameter of interest we are trying to estimate; a constant; e.g. μ
- *Estimator*: the statistic used to estimate the estimand; a random variable; e.g. \bar{Y}
- *Estimate*: a realization of an estimator from an observed dataset; e.g. $\bar{y} = 36.3$

Estimating μ

- Suppose Y_1, \dots, Y_n is a random sample from a distribution with mean μ
- The estimator \bar{Y} is an *unbiased* estimator of μ , i.e.,

$$E(\bar{Y}) = \mu$$

That is, the mean of the sampling distribution of \bar{Y} equals μ , the population parameter of interest

Confidence Interval for μ

- Suppose Y_1, \dots, Y_n is a random sample from a normal distribution with mean μ and variance σ^2
- Then

$$\bar{Y} \sim N(\mu, \sigma^2/n)$$

(Result 4.2 in previous set of slides)

- We can use this to derive a *confidence interval* (CI) for μ

Confidence Interval for μ

- First define z_p such that

$$\Pr[Z \leq z_p] = p$$

for $Z \sim N(0, 1)$; by symmetry, $z_p = -z_{1-p}$

- z_p is the p^{th} quantile of a standard normal distribution

Confidence Interval for μ

$$1 - \alpha = \Pr[-z_{1-\alpha/2} < Z < z_{1-\alpha/2}]$$

$$= \Pr[-z_{1-\alpha/2} < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}]$$

$$= \Pr[-z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{Y} - \mu < z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

$$= \Pr[-\bar{Y} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{Y} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

$$= \Pr[\bar{Y} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}]$$

Confidence Interval for μ

- $100(1 - \alpha)\%$ CI for μ

$$\bar{Y} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

or

$$(\bar{Y} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{Y} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}})$$

- Values of $z_{1-\alpha/2}$

$100(1 - \alpha)\%$	α	$z_{1-\alpha/2}$
90%	0.10	1.645
95%	0.05	1.960
99%	0.01	2.576

CI Interpretation, Comment

- Text (p 86): The probability is $1 - \alpha$ that the interval *straddles* the population mean μ
- If we draw 100 different random samples, on average $100(1 - \alpha)\%$ of them will contain μ
- To decrease the width of CI:
 - increase α , i.e., decrease confidence
 - increase sample size

CI Example

- Example 4.8 of the text: SIDS birthweights
- $n = 78, \bar{Y} = 2994g, \sigma = 800g$
- A 95% CI for the mean birthweight

$$2994 \pm 1.96 \frac{800}{\sqrt{78}} = (2816, 3172)$$

- A 99% CI for the mean birthweight

$$2994 \pm 2.58 \frac{800}{\sqrt{78}} = (2760, 3228)$$

Assumptions

- Y s are sampled from a normal distribution
- Variance is known

But . . .

- What do we do if the variance is unknown?
- If σ^2 is not known, we can estimate it with s^2
- However, the distribution of

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

is not normal

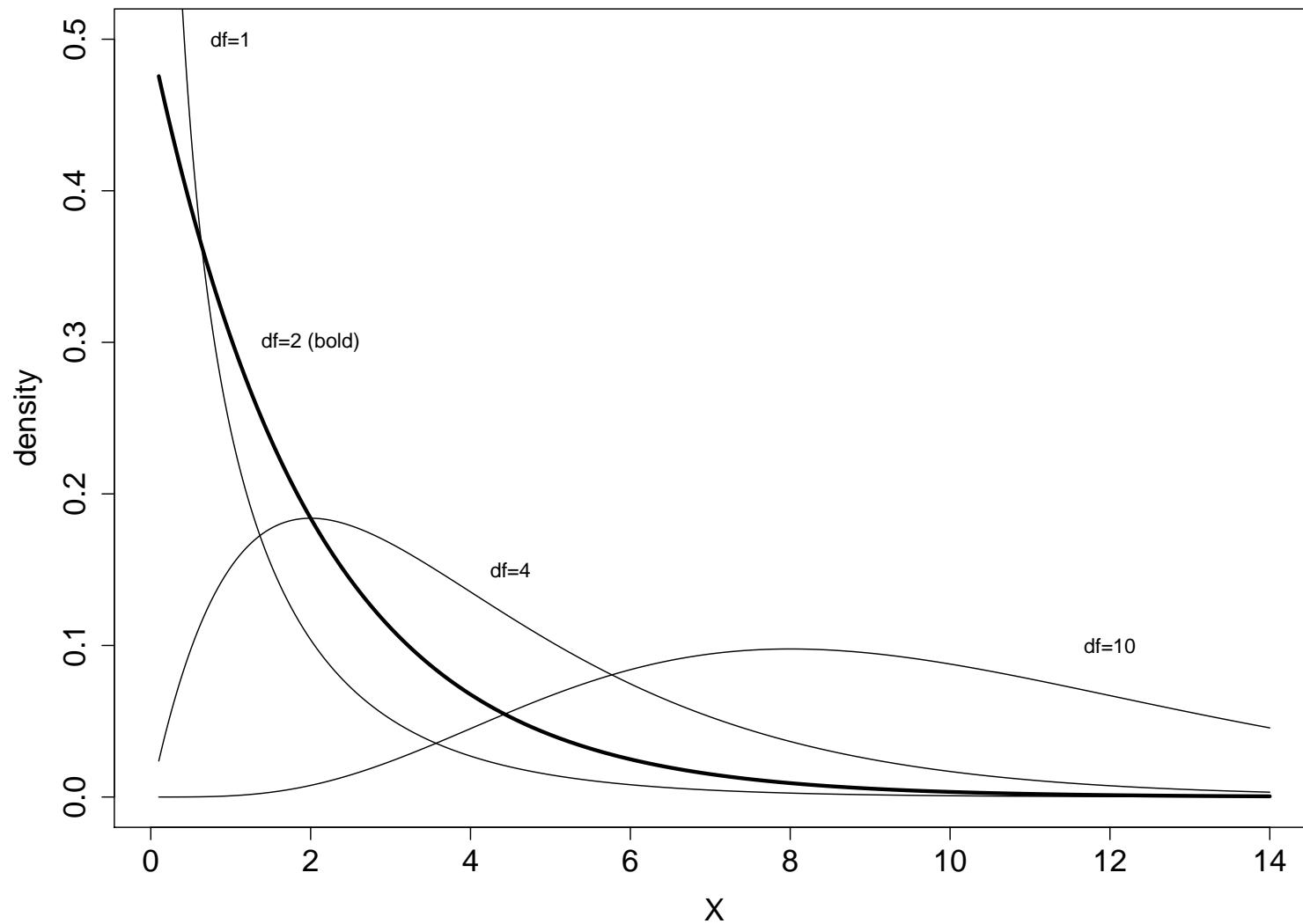
Distribution of s^2

- Result 4.4 (text, p 95): If a random variable Y is normally distributed with mean μ and variance σ^2 , then for a random sample of size n , the quantity

$$\frac{(n - 1)s^2}{\sigma^2}$$

has a chi-square distribution with $n - 1$ degrees of freedom, which we denote by χ_{n-1}^2

χ^2 Distribution



t Distribution

- Let $Z \sim N(0, 1)$ and $W \sim \chi_{\nu}^2$
- If Z and W are independent, then

$$T = \frac{Z}{\sqrt{W/\nu}}$$

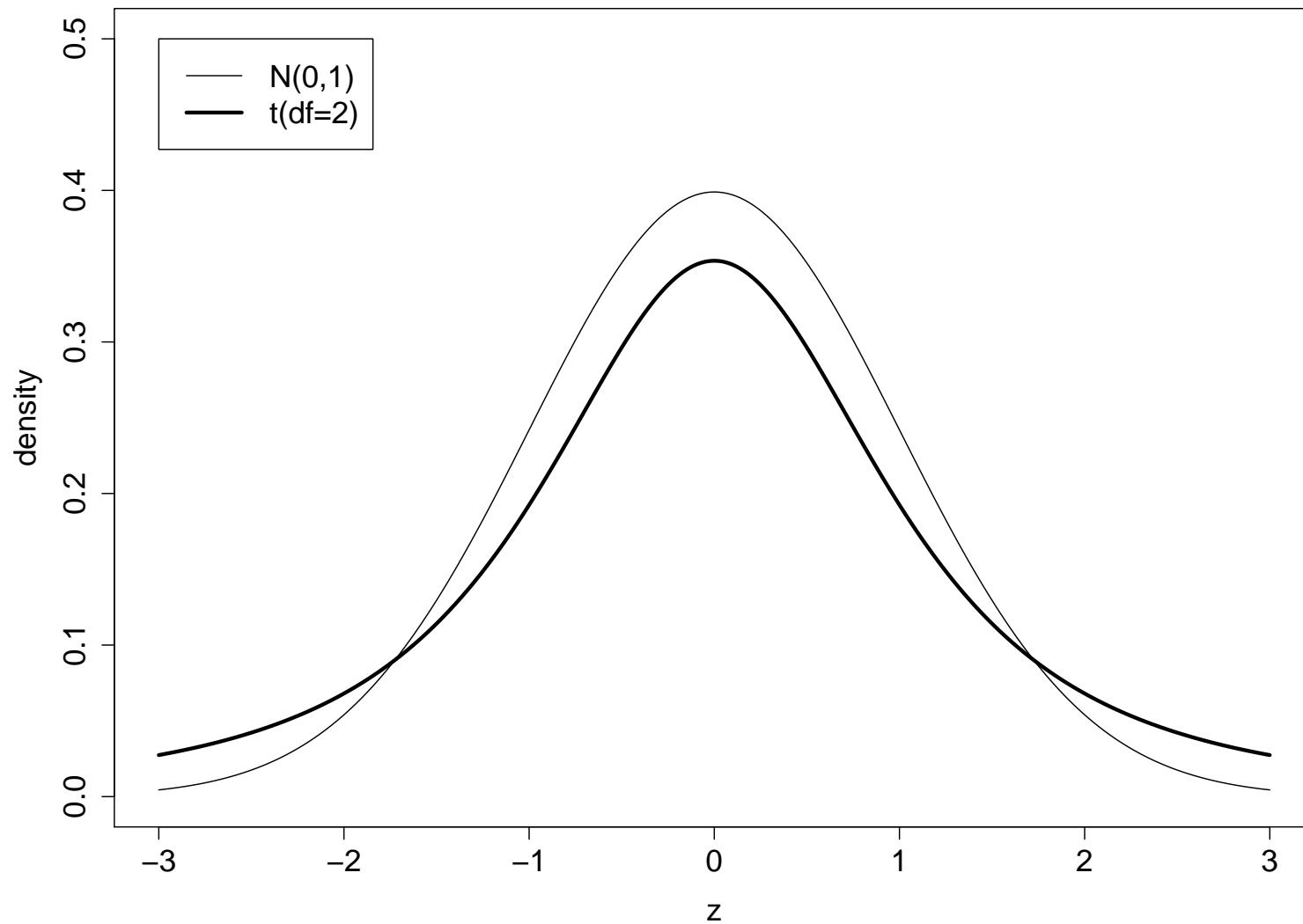
will follow the *t*-distribution with ν degrees of freedom.

- We know

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \text{ and } W = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Can show that \bar{Y} and s^2 are independent

t Distribution



CI for μ when σ^2 unknown

- Substituting, we get

$$T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}(\bar{Y} - \mu)/\sigma}{\sqrt{\{(n-1)s^2/\sigma^2\}/(n-1)}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

- Thus a $100(1 - \alpha)\%$ CI for μ is given by

$$\bar{Y} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$$

- Note 1: We are still assuming the Y s are normal
- Note 2: If $n \geq 30$, we can use z as a reasonable approximation for t

Example

- Now suppose we have the birthweights of a sample of SIDS victims but that the variance is unknown
- $\bar{Y} = 2920.0$, $s = 792.86$ and $n = 23$
- $t_{22,0.975} = 2.07$
- 95% CI for μ :

$$\begin{aligned} 2920.0 \pm 2.07 \left(\frac{792.86}{\sqrt{23}} \right) &= 2920.0 \pm 342.9 \\ &= (2577.1, 3262.8) \end{aligned}$$

- Note that here we multiply the (estimated) s.e. by 2.07 rather than the normal distribution's 1.96 as a penalty for not knowing σ

Quantiles of t

- How to find $t_{22,0.975} = 2.07$?
- Text, Table A.4 page 822: column 4, row 22
- R:

```
> qt(0.975,22)
[1] 2.073873
```

- SAS:

```
data;
x=quantile('T',0.975,22);
proc print;
```

Obs	x
1	2.07387

CIs Using Software

- R:

```
> t.test(x)$conf.int
```

```
[1] 2577.113 3262.830  
attr(,"conf.level")  
[1] 0.95
```

- SAS PROC TTEST (edited output):

```
proc ttest;  
var x;
```

The TTEST Procedure

N	Mean	Std Dev	Std Err	Minimum	Maximum
23	2920.0	792.9	165.3	1252.8	4369.2
Mean	95% CL Mean	Std Dev	95% CL Std Dev		
2920.0	2577.1 3262.8	792.9	613.2 1122.2		

Non-normal Data

- If the Y s are not normally distributed, we use the CLT:
- If Y_1, \dots, Y_n is a random sample from a distribution with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$ for $i = 1, \dots, n$, then \bar{Y} is approximately distributed as $N(\mu, \sigma^2/n)$ for large n
- The use of the CLT to construct a CI for μ requires knowledge of σ^2
- To use the CLT when σ^2 unknown requires *Slutsky's Theorem*

Slutsky's Theorem

- If X_n is a sequence of random variables that converges in distribution to X , and
- Y_n is a sequence of random variables that converges in probability to a constant c ,
- Then $W_n = X_n Y_n$ converges in distribution to cX
- That is

$$\lim_{n \rightarrow \infty} \Pr[W_n \leq w] = \Pr[cX \leq w]$$

Non-normal Data

- Let

$$X_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad \text{and} \quad Y_n = \sqrt{\frac{\sigma^2}{s^2}}$$

- We know $X_n \xrightarrow{d} Z \sim N(0, 1)$ and $\sigma^2/s^2 \xrightarrow{p} 1$
- Then Slutsky's Theorem implies

$$W_n = X_n Y_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{s^2}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

will be approximately $\sim N(0, 1)$

- The approximation gets better as $n \rightarrow \infty$

Large Sample CI for μ

- If n is sufficiently large, an approximate $100(1 - \alpha)\%$ CI for μ is

$$\bar{Y} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

- This is true regardless of the original distribution of the Y s

Example

- A survey was conducted to estimate the mean age that smoking was started among women who smoke.
A random sample of 243 smoking women in NC found $\bar{y} = 16.8$ and $s = 2.36$.
- A 95% CI for the mean age of smoking onset is:

$$16.8 \pm 1.96 \left(\frac{2.36}{\sqrt{243}} \right) = (16.5, 17.1)$$

Summary of CIs for μ

Normal	σ^2 known	n large	Confidence Interval
✓	✓	—	$\bar{Y} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$
—	✓	✓	$\bar{Y} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$
✓	—	—	$\bar{Y} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n})$
—	—	✓	$\bar{Y} \pm z_{1-\alpha/2}(s/\sqrt{n})$
—	—	—	Transform; nonparametrics

Non-normal Data with Small Sample Size

- With a small sample size it is difficult to test for normality
- Transformation of the data
- Nonparametric methods
 - Bootstrap
 - CI for median

Bootstrap t-intervals

- Empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I[X_i \leq x]$$

- Statistical theory indicates $F_n(x) \xrightarrow{p} F(x)$ where F is the population distribution function
- Bootstrap: approximate the sampling distribution of a statistic (in this case the sample mean) by repeatedly sampling (with replacement) from the empirical distribution function F_n
- See text, section 8.10.2

Bootstrap t-intervals

- Bootstrap t-interval: an approximate $100(1 - \alpha)\%$ CI for μ is

$$(\bar{Y} - \hat{t}_{(1-\alpha/2)} \frac{s}{\sqrt{n}}, \quad \bar{Y} - \hat{t}_{(\alpha/2)} \frac{s}{\sqrt{n}})$$

where $\hat{t}_{(1-\alpha/2)}$ and $\hat{t}_{(\alpha/2)}$ are determined from bootstrap samples as described on the next slide

Bootstrap t-intervals

1. Draw a random sample of size n with replacement from $\{x_1, \dots, x_n\}$; call this $\mathbf{x}^*(1)$
 2. Compute $Z^*(1)$ as described on the next slide
 3. Do steps 1 and 2 a total of B times, to obtain $Z^*(1), Z^*(2), \dots, Z^*(B)$
 4. Let $\hat{t}_{(\alpha/2)}$ be the $\alpha/2$ sample quantile of $\{Z^*(1), \dots, Z^*(B)\}$; similarly for $\hat{t}_{(1-\alpha/2)}$
-
- The order of steps 2 and 3 can be interchanged, first obtaining B bootstrap samples $\mathbf{x}^*(1), \dots, \mathbf{x}^*(B)$ and then computing $Z^*(b)$ for each bootstrap sample

Bootstrap t-intervals

- Step 2. For each bootstrap sample compute

$$Z^*(b) = \frac{\bar{x}^*(b) - \bar{x}}{\hat{se}^*(b)}$$

where $\bar{x}^*(b)$ is the mean of $\mathbf{x}^*(b)$, \bar{x} is the mean of the original sample, and $\hat{se}^*(b)$ is the estimated standard error of $\bar{x}^*(b)$, i.e.,

$$\hat{se}^*(b) = \sqrt{\text{Var}\{\mathbf{x}^*(b)\}/n}$$

where $\text{Var}\{\mathbf{x}^*(b)\}$ is the sample variance of the b^{th} bootstrap sample $\mathbf{x}^*(b)$

Bootstrap t-intervals

- Simulation study; 10,000 simulated datasets of size $n = 20$; $B = 500$ bootstrap samples per dataset
- Calculate empirical coverage probabilities for CIs using t and bootstrap t . That is, for what proportion of the 10,000 simulated datasets do the CIs contain the true mean μ
- A “good” method of generating confidence intervals should have empirical coverage close to the claimed confidence (95% in this example)
- Generate the datasets from three different population distributions, each with mean $\mu = 1$

Simulating Using R

```
sim <- function(void){

  cover_tmp=c(0,0)
  n <- 20
  # y <- rnorm(n,1,1)
  # y <- rexp(n,1)  # mean of exp. 1 is 1
  y <- rchisq(n,1)  # mean of chisq k is k

  mean.y <- mean(y)
  var.y <- var(y)

  # CI using t
  lower.y <- mean.y - qt(0.975,n-1)*sqrt(var.y/n)
  upper.y <- mean.y + qt(0.975,n-1)*sqrt(var.y/n)

  if (lower.y <1 & upper.y >1) cover_tmp[1] <- 1
}
```

```

# bootstrap-t interval
boots <- 500
zs <- matrix(0,1,boots)
for (jj in 1:boots){
  ysamp <- sample(y,size=n,replace=T)
  zs[jj] <- (mean(ysamp)-mean.y)/sqrt(var(ysamp)/n)
}
lower.t <- quantile(zs,0.975)
upper.t <- quantile(zs,0.025)

lower.y <- mean.y - lower.t*sqrt(var.y/n)
upper.y <- mean.y - upper.t*sqrt(var.y/n)

if (lower.y <1 & upper.y >1) cover_tmp[2] <- 1

cover_tmp
}

```

```
# run the simulation
nsims <- 10000
set.seed(43567)
cover <- matrix(0,nsims,2)
for (ii in 1:nsims){
    cover[ii,] <- sim(ii)
}
output <-c(mean(cover[,1]),mean(cover[,2]))
print(output)
```

Bootstrap t-intervals

Empirical Coverage Probabilities

	Population n	distribution	Bootstrap t	“Simple” bootstrap
20	N(1,1)		0.949	0.946
	Chi-squared (1 df)		0.891	0.936
	Chi-squared (1 df)*		0.905	0.942
	Exponential(1)		0.922	0.945
25	N(1,1)		0.949	0.947
	Chi-squared (1 df)		0.922	0.944
	Exponential(1)		0.902	0.939

* Using a different random number seed

Bootstrap Notes

- Many types of bootstrap CIs available
- Bootstrap CIs need not be symmetric
- Large sample theoretical justification; empirically small sample performance good
- R: library("boot")

Outline

- Introduction
- CIs for the mean
 - Parametric, large sample
 - Bootstrap
- CI for quantiles
 - Exact
 - Large sample
- CI for variance

Nonparametric CI for the Median

- Suppose X_1, \dots, X_n are iid according to CDF F
- Let $\zeta_{0.5}$ be the population median
- Construct a symmetric $100(1 - \alpha)\%$ CI by finding largest r such that

$$\Pr[X_{(r)} \leq \zeta_{0.5} \leq X_{(n-r+1)}] \geq 1 - \alpha$$

- Sufficient to find largest r such that

$$\Pr[\zeta_{0.5} < X_{(r)}] \leq \alpha/2$$

Bernoulli Random Variable

- Let Y be a Bernoulli random variable
- Y can take on two values, 0 or 1

$$\Pr[Y = 1] = \pi; \quad \Pr[Y = 0] = 1 - \pi$$

$$E(Y) = \pi; \quad \text{Var}(Y) = \pi(1 - \pi)$$

Binomial Random Variable

- Consider a process that produces independent Bernoulli random variables with the same probability of success π
- Let Y count the number of successes in n trials
- $Y \sim \text{Binomial}(n, \pi)$

$$\Pr[Y = y] = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad y = 0, 1, 2, \dots, n$$

$$E(Y) = n\pi$$

$$\text{Var}(y) = n\pi(1 - \pi)$$

Derivation of CI for Median

- CDF

$$\Pr[X_i \leq x] = F(x)$$

- Therefore

$$\begin{aligned}\Pr[x < X_{(r)}] &= 1 - \Pr[X_{(r)} \leq x] \\ &= 1 - \Pr[\text{at least } r \text{ of the } X_i \leq x] \\ &= 1 - \sum_{i=r}^n \binom{n}{i} F(x)^i \{1 - F(x)\}^{n-i} \\ &= \sum_{i=0}^{r-1} \binom{n}{i} F(x)^i \{1 - F(x)\}^{n-i}\end{aligned}$$

Derivation of CI for Median

- CDF of Binomial($n, \pi = F(x)$)

- If $p = 0.5$, then $F(\zeta_p) = 0.5$

- So

$$\Pr[\zeta_{0.5} < X_{(r)}] = \frac{1}{2^n} \sum_{i=0}^{r-1} \binom{n}{i}$$

- Choose largest r such that

$$\frac{1}{2^n} \sum_{i=0}^{r-1} \binom{n}{i} \leq \alpha/2$$

Derivation of CI for Median: Example

- Using $n = 23$
- CDF of $X \sim \text{Binomial}(23, 0.5)$

x	$\Pr[X \leq x]$
0	1.192093e-07
1	2.861023e-06
2	3.302097e-05
3	2.441406e-04
4	1.299739e-03
5	5.311012e-03
6	1.734483e-02
7	4.656982e-02
:	

- Pick $r = 7$

Derivation of CI for Median

- Values of r for 95% CI for Median

n	r
1-5	0
6-8	1
9-11	2
12-14	3
15-16	4
17-19	5
20-22	6
23-24	7
25-27	8
28-29	9
30-32	10
33-34	11

- Cf. page 269-270 of the text

95% CI Example

- For $n = 23$, choose $r = 7$ and then $n - r + 1 = 17$
- Therefore

$$(x_{(7)}, x_{(17)})$$

gives a 95% CI for the median

- This CI makes no assumptions about the distribution of the Y s
- Note:

$$\frac{1}{2^{23}} \sum_{i=7}^{23-7} \binom{23}{i} = 0.9653 \geq 1 - \alpha$$

```
> sum(dbinom(7:16,23,0.5))
[1] 0.9653103
```

SAS Code and Output

```
proc univariate data=beta cipctldf;  
  var base1;  
run;
```

Quantiles (Definition 5)

Quantile	Estimate	95% Confidence Limits		-----Order Statistics-----		
		Distribution Free		LCL Rank	UCL Rank	Coverage
100% Max	298					
99%	298
95%	252	212	298	21	23	58.75
90%	212	202	298	19	23	83.83
75% Q3	192	162	252	13	22	97.35
50% Median	152	106	186	7	17	96.53
25% Q1	100	74	124	2	11	97.35
10%	80	68	92	1	5	83.83
5%	74	68	80	1	3	58.75
1%	68
0% Min	68					

Large Sample CI for the Median

- The above method of finding a $(1 - \alpha)100\%$ CI for the median is *exact*, i.e., the probability the CI contains $\zeta_{0.5}$ is guaranteed to be at least $(1 - \alpha)$
- Now we derive a large sample CI for the median using the CLT
- This will be approximate in that the probability the CI contains $\zeta_{0.5}$ is approximately $(1 - \alpha)$, with the approximation improving as $n \rightarrow \infty$

Large Sample CI for Any Quantile

- In general,

$$\begin{aligned}\Pr[\zeta_p < X_{(r)}] &= \sum_{i=0}^{r-1} \binom{n}{i} F(\zeta_p)^i \{1 - F(\zeta_p)\}^{n-i} \\ &= \sum_{i=0}^{r-1} \binom{n}{i} p^i q^{n-i}\end{aligned}$$

where $q = 1 - p$

- From the CLT, if $Y \sim \text{Binomial}(n, p)$, then

$$\frac{Y - np + 1/2}{\sqrt{npq}} \sim N(0, 1)$$

- The $1/2$ is a *continuity correction* (see text p. 156)

Large Sample CI for Any Quantile

- Thus

$$\begin{aligned}\Pr[\zeta_p < X_{(r)}] &= \Pr[Y \leq r - 1] \\ &\approx \Pr[Z \leq \frac{(r - 1) - np + 1/2}{\sqrt{npq}}] \\ &= \Phi\left(\frac{r - np - 1/2}{\sqrt{npq}}\right)\end{aligned}$$

- The goal is a symmetric $(1 - \alpha)\%$ CI, so we want

$$\alpha/2 = \Pr[\zeta_p < X_{(r)}] = \Phi\left(\frac{r - np - 1/2}{\sqrt{npq}}\right)$$

- That is

$$-z_{1-\alpha/2} = \frac{r - np - 1/2}{\sqrt{npq}}$$

Large Sample CI for Any Quantile

- This implies

$$r = np + \frac{1}{2} - z_{1-\alpha/2}\sqrt{npq}$$

- Similar reasoning yields

$$s = np + \frac{1}{2} + z_{1-\alpha/2}\sqrt{npq}$$

- For $p = 1/2$:

$$r = \frac{n+1}{2} - z_{1-\alpha/2} \frac{\sqrt{n}}{2}$$

$$s = \frac{n+1}{2} + z_{1-\alpha/2} \frac{\sqrt{n}}{2}$$

Large Sample CI for Any Quantile

- Thus a $100(1 - \alpha)\%$ CI for ζ_p is given by
$$(X_{(\lfloor r \rfloor)}, X_{(\lceil s \rceil)})$$
- Note: n large enough ensures $\lfloor r \rfloor, \lceil s \rceil \in \{1, \dots, n\}$

Large Sample CI for Median: Example

- Suppose $n = 100$ and $\alpha = 0.05$

- Then

$$z_{1-\alpha/2} \frac{\sqrt{n}}{2} = 5(1.96) = 9.8$$

- Rounding (using the floor and ceiling functions) yields:

$$50.5 \pm 9.8 \Rightarrow (x_{(40)}, x_{(61)})$$

- Can show $r = 40$ using the exact method

```
> sum(dbinom(40:60,100,1/2))
[1] 0.9647998
> sum(dbinom(41:59,100,1/2))
[1] 0.943112
>
> 2*sum(dbinom(0:39,100,1/2))
[1] 0.0352002
```

CI for Variance

- Suppose Y_1, \dots, Y_n is a random sample from a normal distribution with mean μ and variance σ^2
- Recall (result 4.4 on p. 95 of the text)

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

- Therefore

$$1 - \alpha = \Pr[\chi_{\alpha/2, n-1}^2 \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi_{1-\alpha/2, n-1}^2]$$

- Implying

$$1 - \alpha = \Pr \left[\frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2} \right]$$

CI for Variance

- Because the χ^2 distribution is not symmetric, we need to look up both $\chi_{\alpha/2,n-1}^2$ and $\chi_{1-\alpha/2,n-1}^2$
- This CI is dependent on the Y_s being from a normal distribution

CI for Variance Example

- Using the same data as for the example of the CI for the median (slides on pp. 44, 46 & 47)
- $n = 23; s^2 = 3701.36$
- R: `qchisq(0.025,22)`
SAS: `data; x=quantile('Chisq',0.025,22);`
Table A.3, page 821
- $\chi^2_{0.025,22} = 10.98; \chi^2_{0.975,22} = 36.78$
- Therefore, 95% CI for σ^2 :
$$(22(3701.36)/36.78, 22(3701.36)/10.98)$$
$$= (2213.973, 7416.203)$$
- 95% CI for $\sigma = (47.05, 86.12)$

SAS Code and Output

```
proc univariate data=beta cibasic;  
    var base1;  
run;
```

Basic Confidence Limits Assuming Normality

Parameter	Estimate	95% Confidence Limits	
Mean	150.78261	124.47394	177.09128
Std Deviation	60.83880	47.05242	86.10828
Variance	3701	2214	7415

CI for Variance – Non-normal Data

- Large sample theory

$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} N(0, (\alpha_4 - 1)\sigma^4)$$

where $\alpha_4 = E(X - \mu)^4/\sigma^4$ is the *kurtosis*
(cf. Dudewicz and Mishra, *Modern Mathematical Statistics*, p. 325)

- “Crude approximation”: replace usual CI with

$$\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2,n-1}^2(1+g_2/n)}, \frac{(n-1)s^2}{\chi_{\alpha/2,n-1}^2(1+g_2/n)} \right)$$

where $g_2 = a_4 - 3$ and a_4 is an estimate of α_4
(cf. Solomon and Stephens, *Encyclopedia of Stat Sci*)

CI for Variance – Non-normal Data

- Nonparametric approach such as bootstrap (cf. Efron and Tibshirani, *An Introduction to the Bootstrap*, Ch. 14)
- Software?

BIOS 662 Fall 2018

Tests of Hypotheses

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Basic Approach

1. Set up a hypothesis
 2. Collect data
 3. Infer from the data whether the hypothesis is plausible
- Examples:
 - Is mean BP the same in diabetics and non-diabetics?
 - Will folic acid supplementation reduce the risk of stroke?

Null Hypothesis: H_0

- Null hypothesis H_0 : the hypothesis to be tested

- Example: Null hypothesis for folic acid study

The incidence of stroke is the same in those taking folic acid supplements as in those not taking folic acid supplements

- See Note 4.16 in the text. Typically H_0 is:

- the prevailing view or straw man, or
- the most parsimonious hypothesis

Null and Alternative

- In a test of a hypothesis, we are testing whether some population parameter has a particular value
- For example,

$$H_0 : \theta = \theta_0$$

where θ_0 is a specified constant

- The **alternative hypothesis** is the complement of the null hypothesis

$$H_A : \theta \neq \theta_0$$

Test Statistic

- Once the data are collected, we compute the value of a *test statistic* related to θ , say $S(\hat{\theta})$
- $S(\hat{\theta})$ is a random variable, because it is computed from a sample
- $S(\hat{\theta})$ will have a particular probability distribution under the assumption H_0 , say $F_0(S(\hat{\theta}))$

Test Statistic

- Under F_0 , we compute the probability that we would observe $S(\hat{\theta})$ or a value more extreme than $S(\hat{\theta})$ if the null H_0 is true
- If this probability is large, the data are consistent with H_0
- If this probability is small, there are two possibilities:
 1. An unlikely event has occurred
 2. H_0 is not true

Interpretation

- Usually if the probability is small, we conclude H_0 is not true; i.e., we “reject” H_0
- If the probability is large, we have not proved H_0 . We say that “we failed to reject H_0 ”
- We can never prove H_0 is true!
- Also: we don’t “accept the alternative”

Significance Level

- How do we decide if the probability is too small?
- Prior to seeing the data, we select a value α such that:

If the computed probability is less than or equal to α , we reject H_0

- α is known as the *significance level*

Critical Region and Value

- We have a statistic $S(\hat{\theta})$ with distribution F_0 under the null hypothesis
- We specify α and under F_0 determine a *critical region* or *rejection region* C_α such that

$$\Pr[S(\hat{\theta}) \in C_\alpha | H_0] = \alpha$$

- Values at the boundaries of C_α are called *critical values*

Critical Region and Value

- From the data we compute the value of $S(\hat{\theta})$
- If $S(\hat{\theta}) \in C_\alpha$, we reject H_0
- If $S(\hat{\theta}) \notin C_\alpha$, the data are consistent with H_0 (or, at least, not *inconsistent* with H_0) and we do not reject H_0

Tests of Hypotheses: Seven Steps

1. Design study (sample size depends on steps 2-4)
2. Establish null hypothesis
3. Determine test statistic to be employed
4. Choose significance level α and establish C_α
5. Carry out study and collect data
6. Compute statistic from data
7. If statistic is in C_α , reject H_0

Example

- Does calcium supplementation affect blood pressure in African Americans with high blood pressure?
- Study: Enroll 10 AA men with hypertension; measure their BP; ask them to take calcium tablets for 3 weeks and re-measure their BP
- Aside: Later in the semester we will look at how to determine whether $n = 10$ is a large enough sample size to provide a reasonable test of the hypotheses.

Example cont.

- Let θ denote the mean BP change after 3 weeks
- Hypotheses

$$H_0 : \theta = 0 \text{ vs } H_A : \theta \neq 0$$

- Let $Y_i = \text{BP at 3 weeks} - \text{BP at baseline}$ for the i^{th} individual in the study, $i = 1, \dots, 10$
- $\hat{\theta} = \bar{Y}$

Example cont.

- Intuition: We want to reject H_0 if \bar{Y} is far from $\theta_0 = 0$, i.e., if

$$|\bar{Y}| > c$$

for some constant c

- In particular, we want c such that

$$\Pr [|\bar{Y}| > c \mid H_0] = \alpha$$

Example cont.

- Equivalently, want c such that

$$\Pr \left[\left| \frac{\bar{Y}}{s/\sqrt{n}} \right| > \frac{c}{s/\sqrt{n}} \mid H_0 \right] = \alpha$$

- Assuming Y_i are iid $N(\theta, \sigma^2)$, under H_0 ,

$$\frac{\bar{Y}}{s/\sqrt{n}} \sim t_{n-1}$$

- Thus choose c such that

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

Example cont.

- So we reject H_0 if

$$|\bar{Y}| > c = t_{n-1, 1-\alpha/2} s / \sqrt{n}$$

that is, if

$$\frac{|\bar{Y}|}{s / \sqrt{n}} > t_{n-1, 1-\alpha/2}$$

- Equivalently

$$C_\alpha = \{T : |T| > t_{n-1, 1-\alpha/2}\}$$

where

$$T = \frac{\bar{Y}}{s / \sqrt{n}}$$

Example cont.

- In the calcium supplementation example,

$$S(\hat{\theta}) = S(\bar{Y}) = \frac{\bar{Y} - \theta_0}{s/\sqrt{n}} \sim t_9$$

- If $\alpha = 0.05$, the critical region is

$$C_{0.05} = \{T : |T| > t_{9,0.975} = 2.26\}$$

where

$$T = \frac{\bar{Y} - 0}{s/\sqrt{10}}$$

Example cont.

Calcium supplementation in African-American men

ID	treatment	before	after	change
1.	calcium	107	100	-7
2.	calcium	110	114	4
3.	calcium	123	105	-18
4.	calcium	129	112	-17
5.	calcium	112	115	3
6.	calcium	111	116	5
7.	calcium	107	106	-1
8.	calcium	112	102	-10
9.	calcium	136	125	-11
10.	calcium	102	104	2

Example using R

```
> x <- c(-7,4,-18,-17,3,5,-1,-10,-11,2)
> se <- sd(x)/sqrt(length(x))
> mean(x)/se
[1] -1.808411

> t.test(x)
```

One Sample t-test

```
data: x
t = -1.8084, df = 9, p-value = 0.104
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
-11.254545 1.254545
```

Example using SAS PROC UNIVARIATE

```
proc univariate;  
  var x;
```

The UNIVARIATE Procedure

Variable: x

Moments

N	10	Sum Weights	10
Mean	-5	Sum Observations	-50
Std Deviation	8.74325137	Variance	76.44444444

Tests for Location: Mu0=0

Test	-Statistic-	-----p Value-----
Student's t	t -1.80841	Pr > t 0.1040

Example using SAS PROC TTEST

```
proc ttest;  
  var x;
```

The TTEST Procedure

Variable: x

N	Mean	Std Dev	Std Err	95% CL Mean
10	-5.0000	8.7433	2.7649	-11.2545 1.2545

DF	t Value	Pr > t
9	-1.81	0.1040

Example cont.

- Because the observed $t = -1.8084$ is not in the critical region, we do not reject H_0
- The data are consistent with no effect of calcium on BP

Types of Errors

		Nature / Truth	
		H_0 true	H_A true
Decision	Do not reject H_0	✓	Type II
	Reject H_0	Type I	✓

- Type I error: Reject H_0 when H_0 true
that is, false positive
- Type II error: Do not reject H_0 when H_A true
that is, false negative

Types of Errors

- Type I error

$$\alpha = \Pr \left[S(\hat{\theta}) \in C_\alpha \mid H_0 \right]$$

- Type II error

$$\beta = \Pr \left[S(\hat{\theta}) \notin C_\alpha \mid H_A \right]$$

- Power

$$1 - \beta = \Pr \left[S(\hat{\theta}) \in C_\alpha \mid H_A \right]$$

that is, the probability of rejecting H_0 when H_A is true

Power

- Recall $H_A : \theta \neq \theta_0$
- Power: $\Pr [S(\hat{\theta}) \in C_\alpha \mid H_A]$
- Power depends on the value of θ

$$\Pr [S(\hat{\theta}) \in C_\alpha \mid \theta] \equiv P(\theta)$$

- Note

$$P(\theta_0) = \alpha$$

Alternative Hypotheses

- Different possible alternatives

$$H_A : \theta \neq \theta_0 \text{ (two-sided)}$$

$$H_A : \theta > \theta_0 \text{ (one-sided)}$$

$$H_A : \theta < \theta_0 \text{ (one-sided)}$$

Alternative Hypotheses

- In the calcium supplementation example,

$$H_A : \theta \neq 0; \quad C_{0.05} = \{T : |T| > 2.26\}$$

- If we took a 1-sided alternative,

$$H_A : \theta < 0; \quad C_{0.05} = \{T : T < -1.83\}$$

- Because $T = -1.8084$, we do not reject H_0 in either case

Alternative Hypotheses

- 2-sided test addresses: Does calcium **change** BP?
- 1-sided test addresses: Does calcium **lower** BP?
- If we applied this 1-sided test and obtained $T = 2.5$, we would not reject H_0 because our test did not ask if calcium raised or changed BP
- We must choose the alternative hypothesis **before seeing the data**
- Friedman et al. (p. 98) “In general, two-sided tests should be used unless there is strong justification for expecting a difference in only one direction”

P-value

- Definition 4.24. The *p-value* is the smallest significance level α for which the observed data indicate the null hypothesis should be rejected
- Probability of obtaining test statistic as unlikely or more unlikely than the observed test statistic if the null hypothesis is true

Calcium Example Revisited

- Recall $T \sim t_9$; $t_{9,0.975} = 2.26$; $t_{9,0.95} = 1.83$
- p-value for 2-sided test = $2 \Pr [T < -1.8084] = 0.1040$
- p-value for 1-sided test = $\Pr [T < -1.8084] = 0.0520$
- R

```
> 2*pt(-1.8084,9)
[1] 0.1039981
```

```
> pt(-1.8084,9)
[1] 0.05199907
```

BIOS 662 Fall 2018

One Sample Tests for Location

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Small sample, normally distributed
- Large sample
- Nonparametric
 - Sign test
 - Wilcoxon signed rank test

Applications

- Cross-sectional study: Collect data to test a hypothesis about the mean or median of Y
- Paired data; examples:
 - Study of how a characteristic changes from before to after a treatment
 - Study of twins
 - Individually-matched case-control study

Small Sample, Normal Distribution

- Consider a small sample, Y_1, \dots, Y_n , iid with $Y_i \sim N(\mu, \sigma^2)$,
- From the previous set of notes, test statistic:

$$T = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

- For two-sided alternative $H_A : \mu \neq \mu_0$, critical region:

$$C_\alpha = \{t : |t| > t_{n-1, 1-\alpha/2}\}$$

- For one-sided alternative $H_A : \mu > \mu_0$, critical region:

$$C_\alpha = \{t : t > t_{n-1, 1-\alpha}\}$$

Small Sample, Normal Distribution

- P-value: Probability of obtaining a test statistic as extreme or more extreme than the one observed from the sample
- For two-sided alternative $H_A : \mu \neq \mu_0$,

$$p = \Pr[T \leq -|t|] + \Pr[T \geq |t|]$$

where $T \sim t_{n-1}$. Equivalently

$$p = 2 \Pr[T \leq -|t|] = 2 \Pr[T \geq |t|]$$

- For one-sided alternative $H_A : \mu > \mu_0$,

$$p = \Pr[T \geq t]$$

SIDS Example

- Example: Text page 281; problem 8.2
- Investigators are interested in whether babies that die of SIDS have different birthweight than babies who do not die of SIDS.
- A study of 22 dizygotic twins compared the birthweight of the baby who died with the baby who did not die.
- Let our random variable Y be the weight of the SIDS baby (twin) minus the weight of non-SIDS baby (twin)

$$H_0 : \mu_{\text{diff}} = 0$$

$$H_A : \mu_{\text{diff}} \neq 0$$

SIDS Example cont.

- Critical region at $\alpha = 0.05$

$$C_{0.05} = \{t : |t| > t_{21,0.975} = 2.08\}$$

- $\bar{y} = 0.1818$, $s = 369.57$, $n = 22$

$$t = \frac{\bar{y}}{s/\sqrt{n}} = 0.0023$$

- P-value

$$p = 2 \times \Pr[T \leq -0.0023] = 0.9982$$

Example: Using R

```
> t.test(sids.diffs)
```

One Sample t-test

```
data: sids.diffs  
t = 0.0023, df = 21, p-value = 0.9982  
alternative hypothesis: true mean is not equal to 0
```

```
> t.test(sids.diffs,alternative="less")
```

One Sample t-test

```
data: sids.diffs  
t = 0.0023, df = 21, p-value = 0.5009  
alternative hypothesis: true mean is less than 0
```

Example: Using SAS

```
proc ttest;  
  var diff;
```

Statistics

Variable	N	Lower CL		Upper CL	
		Mean	Mean	Mean	Mean
diff	22	-163.7	0.1818	164.04	

T-Tests

Variable	DF	t Value	Pr > t
diff	21	0.00	0.9982

Small Sample

- t test assumptions:
 - Observations are independent
 - Sample is from the normal distribution

Large Sample

- For large sample, using the normal approximation (CLT)

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

and Slutsky's theorem

$$Z = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \text{ is approximately } N(0, 1)$$

- Approximation improves as $n \rightarrow \infty$
- Note: Y s do not need to be normally distributed

Large Sample Example

- Example: Iron deficiency
- Iron deficiency anemia is an important nutritional health issue.
- A study was conducted of 63 boys aged 9-11 from families with income below the poverty level.
- The mean daily iron intake in the U.S. population is known to be 14.45 mg.
- Question: Is iron intake in boys associated with family income?

Large Sample Example cont.

- Conduct large sample test

$$H_0 : \mu = 14.45; \quad H_A : \mu \neq 14.45$$

$$C_{0.05} = \{z : |z| > 1.96\}$$

$$\bar{y} = 12.5; \quad s^2 = 22.5625; \quad n = 63$$

$$z = \frac{12.5 - 14.45}{\sqrt{22.5626/63}} = -3.26$$

- Reject H_0
- Question: What is $2\Phi(-3.26)$?

```
> 2*pnorm(-3.26)
[1] 0.001114122
```

Testing/Estimation: Large Sample

- Testing $H_0 : \mu = \mu_0$ versus $H_A : \mu \neq \mu_0$

$$C_\alpha = \{z : |z| > z_{1-\alpha/2}\}$$

where

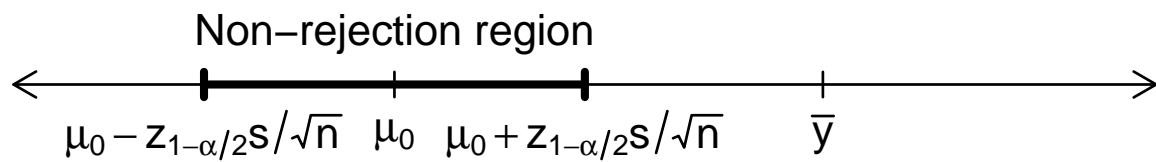
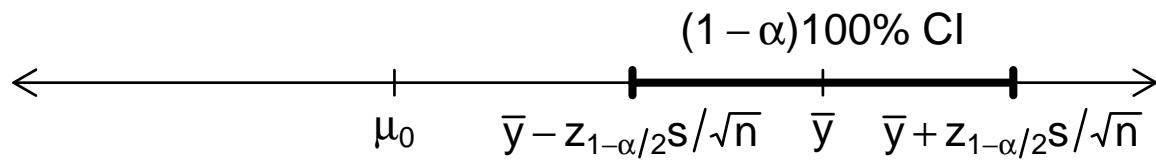
$$z = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

- Estimation: confidence interval for μ

$$\bar{y} \pm z_{1-\alpha/2} \frac{s}{\sqrt{n}}$$

- Can show: Reject H_0 if and only if the CI excludes μ_0

Testing/Estimation: Large Sample



Testing/Estimation: Large Sample

- Theorem: Reject H_0 if and only if CI excludes μ_0
- Sketch of proof: Suppose CI excludes μ_0 , that is,

$$\mu_0 \notin [\bar{y} - z_{1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2} \frac{s}{\sqrt{n}}]$$

Without loss of generality, assume

$$\mu_0 < \bar{y} - z_{1-\alpha/2} \frac{s}{\sqrt{n}},$$

This implies

$$z_{1-\alpha/2} < \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \equiv z$$

implying $z \in C_\alpha$

Testing/Estimation

- Equivalent; two sides of the same coin
- See text Section 4.7
- Emphasize testing beforehand (to inform power, sample size calculations) and estimation afterward
- Or, test first and then use the point estimate and confidence interval to describe the results

Small Sample, Non-normal

- Transformation, bootstrap
- Nonparametric tests:
 - Sign test
 - Wilcoxon signed rank test
- Read text sections 8.1–8.5

Sign Test

- Suppose Y_1, \dots, Y_n are iid continuous from F with median $\zeta_{0.5}$
- Hypotheses

$$H_0 : \zeta_{0.5} = \zeta_{0.5,0} \text{ (median = specified value)}$$

$$H_A : \zeta_{0.5} \neq \zeta_{0.5,0}$$

- If the true median is $\zeta_{0.5,0}$ and F is continuous,

$$\Pr[Y < \zeta_{0.5,0}] = \Pr[Y > \zeta_{0.5,0}] = 0.5$$

for any randomly selected observation Y

Sign Test

- Let R be the number of observations $> \zeta_{0.5,0}$
- Under H_0

$$R \sim \text{Binomial}(n, 0.5)$$

$$\begin{aligned}\Pr[R \leq r] &= \sum_{i=0}^r \binom{n}{i} \left(\frac{1}{2}\right)^i \left(1 - \frac{1}{2}\right)^{n-i} \\ &= \frac{1}{2^n} \sum_{i=0}^r \binom{n}{i}\end{aligned}$$

Sign Test

- Critical region

$$C_\alpha = \{r : r \leq r_{\alpha/2} \text{ or } r \geq r_{1-\alpha/2}\}$$

where $r_{\alpha/2}$ and $r_{1-\alpha/2}$ are such that

$$\Pr[R \leq r_{\alpha/2} \mid H_0] + \Pr[R \geq r_{1-\alpha/2} \mid H_0] \leq \alpha$$

Sign Test

- Because $\text{Binomial}(n, 0.5)$ is symmetric,

$$r_{1-\alpha/2} = n - r_{\alpha/2}$$

- Thus need $r_{\alpha/2}$ such that

$$\Pr[R \leq r_{\alpha/2} \mid H_0] = \frac{1}{2^n} \sum_{i=0}^{r_{\alpha/2}} \binom{n}{i} \leq \frac{\alpha}{2}$$

- Choose largest $r_{\alpha/2}$ such that this inequality holds
- For $n = 10$, the critical region for a 2-sided test with $\alpha = 0.05$ is $C_{0.05} = \{0, 1, 9, 10\}$

Sign Test

- CDF for Binomial($n = 10, \pi = 0.5$)

Cumulative		
r	Probability	$2 \cdot \Pr(R \leq r)$
0	0.0010	0.0020
1	0.0107	0.0214
2	0.0547	0.1094
3	0.1719	0.3437
4	0.3770	0.7539
5	0.6231	
6	0.8282	
7	0.9453	
8	0.9893	
9	0.9990	
10	1.0000	

Sign Test Example

Calcium supplementation in African-American men

	treatment	before	after	diff
1.	calcium	107	100	-7
2.	calcium	110	114	4
3.	calcium	123	105	-18
4.	calcium	129	112	-17
5.	calcium	112	115	3
6.	calcium	111	116	5
7.	calcium	107	106	-1
8.	calcium	112	102	-10
9.	calcium	136	125	-11
10.	calcium	102	104	2

Sign Test Example cont.

- Testing

$$H_0 : \zeta_{0.5} = 0 \text{ vs. } H_A : \zeta_{0.5} \neq 0$$

- For $n = 10$ and $\alpha = 0.05$, $C_{0.05} = \{0, 1, 9, 10\}$
- $r = 4$ is not in $C_{0.05}$
- Hence do not reject H_0
- P-value

$$2 \times \left\{ \frac{1}{2^{10}} \sum_{i=0}^4 \binom{10}{i} \right\} = 0.754$$

$$= 1 - \frac{1}{2^{10}} \binom{10}{5} = 1 - 0.246$$

Sign Test Example cont.

- R code

```
> 2*sum(dbinom(0:4,10,0.5))  
[1] 0.7539063
```

```
> 2*pbinom(4,10,0.5)  
[1] 0.7539063
```

```
> # First need to install the package BSDA  
> # Basic Statistics and Data Analysis  
> library("BSDA")  
> SIGN.test(diff)
```

One-sample Sign-Test

```
data: diff  
s = 4, p-value = 0.7539  
alternative hypothesis: true median is not equal to 0
```

Sign Test Example cont.

- SAS proc univariate output:

The UNIVARIATE Procedure

Variable: diff

Moments

N	10	Sum Weights	10
Mean	-5	Sum Observations	-50
Std Deviation	8.74325137	Variance	76.44444444
Skewness	-0.3378852	Kurtosis	-1.5550482
Uncorrected SS	938	Corrected SS	688
Coeff Variation	-174.86503	Std Error Mean	2.76485885

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value-----	
Student's t	t	-1.80841	Pr > t	0.1040
Sign	M	-1	Pr >= M	0.7539
Signed Rank	S	-13.5	Pr >= S	0.1934

Sign Test: Comments

- Typically $\zeta_{0.5,0} = 0$
- Alternative formulation of the null

$$\Pr[Y < \zeta_{0.5,0}] = \Pr[Y > \zeta_{0.5,0}]$$

- For example, if comparing difference in outcome for a new drug versus control, the null says the probability the new drug is better than the control is the same as the probability the new drug is worse than the control
- Delete any observations $= \zeta_{0.5,0}$ and reduce n by 1 for each

Sign Test: Large Samples

- If n is large, we can use a version of the CLT for the sign test
- Recall, for $R \sim \text{Binomial}(n, \pi)$

$$E(R) = n\pi, \quad \text{Var}(R) = n\pi(1 - \pi)$$

- Thus

$$Z = \frac{R - n\pi}{\sqrt{n\pi(1 - \pi)}}$$

will be approximately $N(0, 1)$

- The approximation gets better as $n \rightarrow \infty$

Sign Test: Large Samples

- For the sign test, $H_0 : \pi = 0.5$
- Therefore we compute

$$Z = \frac{R - n/2}{\sqrt{n/4}}$$

- Critical region comes from $\Phi(z)$

Sign Test: Large Samples

- The normal approximation to the sign test works well for $n \geq 40$
- For $40 \leq n \leq 100$, the approximation is better using the following adjustment:

$$Z = \begin{cases} \frac{R-(n+1)/2}{\sqrt{n/4}} & \text{if } (R - n/2) > 1/2 \\ \frac{R-n/2}{\sqrt{n/4}} & \text{if } |R - n/2| \leq 1/2 \\ \frac{R-(n-1)/2}{\sqrt{n/4}} & \text{if } (R - n/2) < -1/2 \end{cases}$$

Sign Test Example

- A study was conducted to compare an automated machine for measuring blood pressure with measures made by a nurse using a standard mercury sphygmomanometer
- 100 people had their blood pressure measured using both techniques
- We use $\text{sign}(Y = \text{BP}_{\text{auto}} - \text{BP}_{\text{nurse}})$

$$H_0 : \Pr[Y < 0] = \Pr[Y > 0]$$

VS.

$$H_A : \Pr[Y < 0] \neq \Pr[Y > 0]$$

Sign Test Example cont.

- For the study:

$$r = 64$$

\Rightarrow

$$r - n/2 > 1/2$$

\Rightarrow

$$z = 13.5/5 = 2.7$$

- Reject H_0 and conclude that the automated machine is more likely to give a higher reading

Binomial Continuity Correction

- Suppose $R \sim \text{Binomial}(n, \pi)$ and

$$X \sim N(n\pi, n\pi(1 - \pi))$$

- By the CLT

$$\Pr[R \leq x] \approx \Pr[X \leq x]$$

- Continuity correction

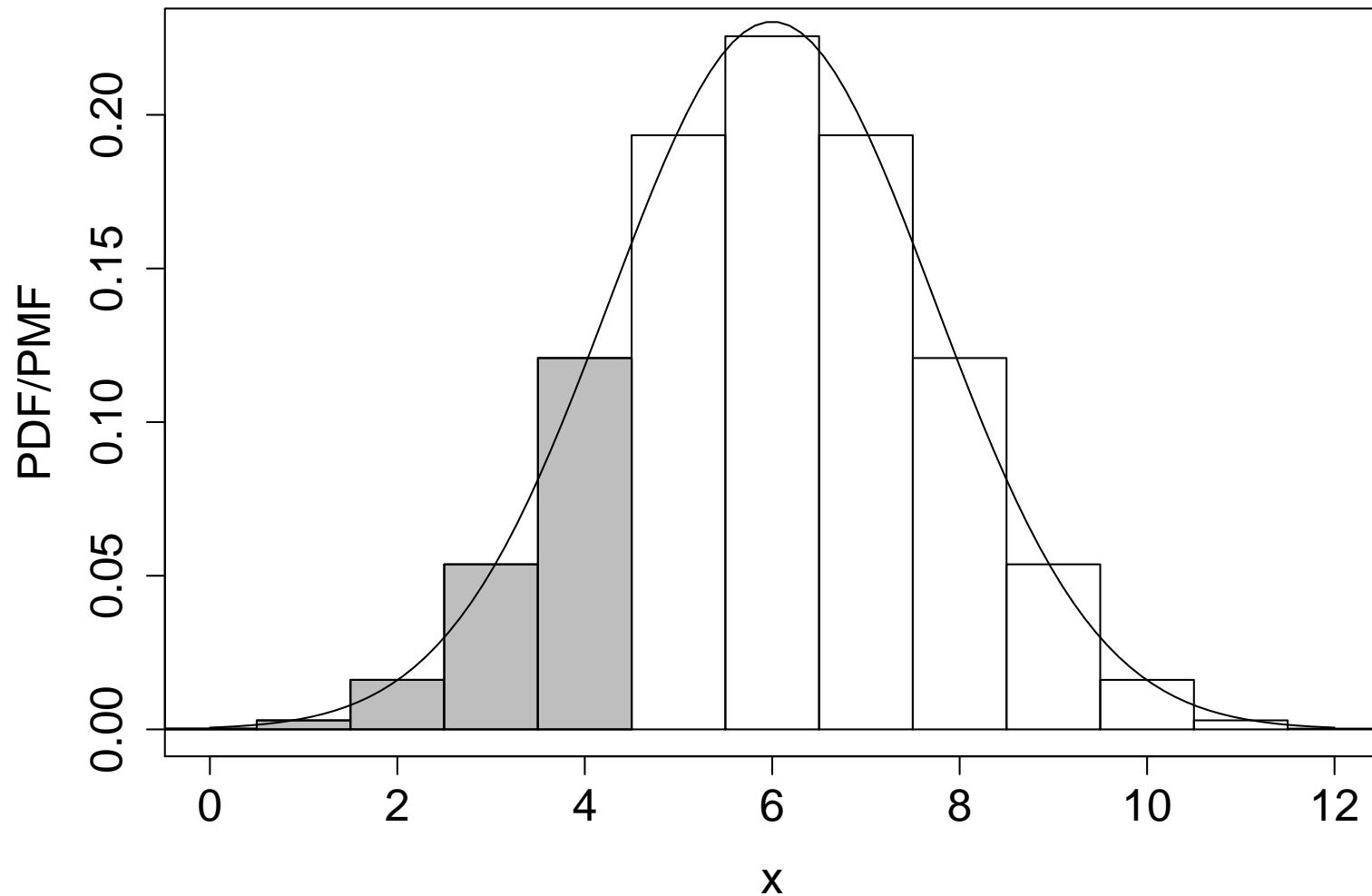
$$\Pr[R \leq x] \approx \Pr[X \leq x + 1/2]$$

- For example, if $n = 12$, $\pi = 0.5$, and $x = 4$

$$0.194 = \Pr[R \leq 4] \approx \Pr[X \leq 4 + 1/2] = 0.193$$

whereas $\Pr[X \leq 4] = 0.124$

Binomial Continuity Correction



Binomial Continuity Correction

- Likewise

$$\Pr[R \geq x] \approx \Pr[X \geq x - 1/2]$$

- Returning to the sign test where $\pi = 1/2$, if $r < \frac{n}{2} - \frac{1}{2}$, then the one-sided p-value will be equal to

$$\begin{aligned}\Pr[R \leq r] &\approx \Pr[X \leq r + 1/2] = \Pr\left[Z \leq \frac{r + 1/2 - n/2}{\sqrt{n/4}}\right] \\ &= \Pr\left[Z \leq \frac{r - (n - 1)/2}{\sqrt{n/4}}\right]\end{aligned}$$

Sign Test: Efficiency

- Definition 8.6. The *relative efficiency* of statistical procedure A compared to B is the ratio of the sample size needed for B to that of A in order for both procedures to have the same statistical power
- If sampling from a normal distribution
 - n small: the sign test is almost as efficient as the t -test
 - As n gets larger, the sign test becomes less efficient and the asymptotic relative efficiency (ARE) is $2/\pi = 0.64$
- If sampling from a non-normal distribution, the sign test can be more efficient

Wilcoxon Signed Rank Test

- Suppose Y_1, Y_2, \dots, Y_n are iid according to a symmetric distribution F with median $\zeta_{0.5}$
- Hypotheses

$$H_0 : \zeta_{0.5} = \zeta_{0.5,0}$$

vs.

$$H_A : \zeta_{0.5} \neq \zeta_{0.5,0}$$

Wilcoxon Signed Rank Test

- Delete any Y_i equal to $\zeta_{0.5,0}$ and adjust n
- Compute $Y'_i = Y_i - \zeta_{0.5,0}$
- Rank the $|Y'_i|$ from smallest to largest
- The statistic S^+ is the sum of the ranks of the observations with Y'_i positive
- S^- defined similarly
- Aside: SAS uses $S^+ - n(n + 1)/4$

Wilcoxon Signed Rank Test Example

Example: Calcium supplementation in African-American men

	treatment	before	after	diff	absol	Y_i	$ Y_i $	rank	sign*rank
1.	calcium	107	100	-7	7	7	7	6	-6
2.	calcium	110	114	4	4	4	4	4	4
3.	calcium	123	105	-18	18	-18	18	10	-10
4.	calcium	129	112	-17	17	-17	17	9	-9
5.	calcium	112	115	3	3	3	3	3	3
6.	calcium	111	116	5	5	5	5	5	5
7.	calcium	107	106	-1	1	-1	1	1	-1
8.	calcium	112	102	-10	10	-10	10	7	-7
9.	calcium	136	125	-11	11	-11	11	8	-8
10.	calcium	102	104	2	2	2	2	2	2

Wilcoxon Signed Rank Test Example cont.

- Table on the course's Sakai site gives critical values
- For $n = 10$, $C_{0.05} = \{S : S \leq 8\}$
- $S^+ = 4 + 3 + 5 + 2 = 14$; $S^- = 41$
- $S = \min\{S^+, S^-\} = 14$
- Therefore, do not reject H_0 : median is 0
- R code:

```
> x <- c(-7, 4, -18, -17, 3, 5, -1, -10, -11, 2)
> wilcox.test(x)
```

```
Wilcoxon signed rank test

data: x
V = 14, p-value = 0.1934
alternative hypothesis: true mu is not equal to 0
```

Wilcoxon Signed Rank Test

- Because

$$\sum_{i=1}^n i = n \left(\frac{n+1}{2} \right)$$

- It follows that

$$S^+ + S^- = n \left(\frac{n+1}{2} \right)$$

- So only smaller values are tabulated

Wilcoxon Signed Rank Test

- How to compute null distribution of signed-rank test?
- Under the null, each ranked observation has probability $1/2$ of having positive sign
- The n signs are independent
- There are 2^n possible outcomes
- Thus each outcome occurs with probability $1/2^n$

Distribution of S^+ Under H_0

- Calculating the null distribution for $n = 4$; a + in the column indicates that the sign of the rank is positive

ranks				S^+
1	2	3	4	
-	-	-	-	0
+	-	-	-	1
-	+	-	-	2
-	-	+	-	3
-	-	-	+	4
+	+	-	-	3
+	-	+	-	4
+	-	-	+	5
-	+	+	-	5
-	+	-	+	6
-	-	+	+	7
+	+	+	-	6
+	+	-	+	7
+	-	+	+	8
-	+	+	+	9
+	+	+	+	10

Distribution of S^+ Under H_0

k	$\Pr[S^+ = k]$	$\Pr[S^+ \leq k]$
0	$1/16$	$1/16 = 0.0625$
1	$1/16$	$1/8 = 0.1250$
2	$1/16$	$3/16 = 0.1875$
3	$2/16$	$5/16 = 0.3125$
4	$2/16$	$7/16 = 0.4375$
5	$2/16$	$9/16 = 0.5625$
6	$2/16$	$11/16 = 0.6875$
7	$2/16$	$13/16 = 0.8125$
8	$1/16$	$7/8 = 0.8750$
9	$1/16$	$15/16 = 0.9375$
10	$1/16$	1

Table A.9: Distribution of S^+ Under H_0

n	k	$\Pr[S^+ \leq k]$
4	0	$1/16 = 0.0625$
	1	$1/8 = 0.125$
	\vdots	\vdots
6	0	$1/2^6 = 0.015625$
	1	$1/32 = 0.03125$
	\vdots	\vdots
7	0	$1/2^7 = 0.0078125$
	1	$2/2^7 = 0.015625$
	2	$3/2^7 = 0.0234$
	3	$5/2^7 = 0.039$

- Bold face values denote critical values listed in Table A.9 of the text on page 834 for one sided $\alpha = 0.025$ and two-sided $\alpha = 0.05$. Is the critical value in the critical region?

Distribution of S^+ Under H_0

- Large sample distribution
- Can show

$$E(S^+) = \frac{n(n+1)}{4} \quad \text{and} \quad \text{Var}(S^+) = \frac{n(n+1)(2n+1)}{24}$$

- If $n \geq 15$,

$$Z = \frac{S^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \sim N(0, 1)$$

Wilcoxon Signed Rank Test: Ties

- If there are two or more observations with the same value of $|Y'|$, the observations are said to be *tied*
- For tied observations we assign the average rank or *midrank*
- Example: $\mathbf{Y} = \{23, 25, 45, 13, 23, 46\}$
Midranks: $\{2.5, 4, 5, 1, 2.5, 6\}$

Wilcoxon Signed Rank Test: Ties

- Can show

$$E(S^+) = \frac{n(n + 1)}{4}$$

- To accommodate ties, the variance is adjusted

$$\text{Var}(S^+) = \frac{n(n + 1)(2n + 1) - \frac{1}{2} \sum_{i=1}^q t_i(t_i - 1)(t_i + 1)}{24}$$

where q equals the number of sets of ties and t_i is the number of observations in the i th set

- For example on previous slide, $q = 1$ and $t_1 = 2$ such that

$$\text{Var}(S^+) = \frac{6(6 + 1)(2 \cdot 6 + 1) - \frac{1}{2} \cdot 2 \cdot 1 \cdot 3}{24}$$

Wilcoxon Signed Rank Test: Ties

- If n is large, use the variance adjusted for ties in the normal approximation
- If n is small and there are ties, need to compute the null distribution from permutation principles, i.e., tables of critical values are not guaranteed to be correct in the presence of ties

Signed Rank Test: Example with Ties

- From Table 8.7, page 281

	SIDS	nonSIDS	Y'	Y'	rank	sgnrnk
1.	1474	2098	-624	624	21	-21
2.	3657	3119	538	538	19	19
3.	3005	3515	-510	510	18	-18
4.	2041	2126	-85	85	3	-3
5.	2325	2211	114	114	4	4
6.	2296	2750	-454	454	15	-15
7.	3430	3402	28	28	1	1
8.	3515	3232	283	283	9	9
9.	1956	1701	255	255	8	8
10.	2098	2410	-312	312	11	-11
11.	3204	2892	312	312	11	11
12.	2381	2608	-227	227	7	-7
13.	2892	2693	199	199	6	6
14.	2920	3232	-312	312	11	-11
15.	3005	3005	0	0		
16.	2268	2325	-57	57	2	-2
17.	3260	3686	-426	426	14	-14
18.	3260	2778	482	482	16.5	16.5
19.	2155	2552	-397	397	13	-13
20.	2835	2693	142	142	5	5
21.	2466	1899	567	567	20	20
22.	3232	3714	-482	482	16.5	-16.5

Signed Rank Test: Example with Ties cont.

- $S^+ = 99.5$; $E(S^+) = 21(22)/4 = 115.5$

- $q = 2$; $t_1 = 3$; $t_2 = 2$ so that

$$\text{Var}(S^+) = \frac{21(22)(43) - \frac{1}{2}[3(2)(4) + 2(1)(3)]}{24} = 827.15$$

- Thus

$$Z = \frac{99.5 - 115.5}{\sqrt{827.15}} = -0.5563$$

- Yielding $p = 2 \times \Phi(-0.5563) = 0.578$

Signed Rank Test: Example with Ties cont.

- R code:

```
> wilcox.test(sids.diffs,exact=F,correct=F)
```

Wilcoxon signed rank test

data: sids.diffs

V = 99.5, p-value = 0.578

alternative hypothesis: true mu is not equal to 0

- SAS uses a slightly different large sample approximation; proc univariate uses the exact version of the test when $n \leq 20$ and a large sample approximation when $n > 20$; here the large sample approximation yields $p = 0.5905$

Efficiency of Signed Rank Test

- If the Y s come from a normal distribution, the ARE of the signed rank test compared to the normal distribution test is $3/\pi = 0.955$
(cf. Lehmann 1998, page 80)
- Gain in efficiency over the sign test is because of an additional assumption: symmetry

Signed rank test: H_A

- Suppose $n = 10$, $\alpha = 0.05$, $H_0 : \zeta_{0.5} = 0$
- For $H_A : \zeta_{0.5} \neq 0$, $S = \min\{S^+, S^-\}$

$$C_{0.05} = \{s : s \leq 8\}$$

- Equivalently

$$C_{0.05} = \{s^+ : s^+ \leq 8 \text{ or } s^+ \geq 47\}$$

or

$$C_{0.05} = \{s^- : s^- \leq 8 \text{ or } s^- \geq 47\}$$

Signed rank test: H_A

- For $H_A : \zeta_{0.5} < 0$, what is the rejection region?

$$C_{0.05} = \{s^+ : s^+ \leq 10\}$$

or

$$C_{0.05} = \{s^- : s^- \geq 45\}$$

BIOS 662 Fall 2018

Two Sample Tests

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Two Sample Test Settings

- Single cross-sectional sample, comparing two sub-samples
- Comparing samples from two different populations (two cross-sectional samples, unmatched case control study)
- Single sample; subjects randomly allocated to different interventions (experiment, clinical trials)

Fundamentals

- Definition 5.2. Two random variables Y_1 and Y_2 are *independent* if for all y_1 and y_2

$$\Pr[Y_1 \leq y_1, Y_2 \leq y_2] = \Pr[Y_1 \leq y_1] \Pr[Y_2 \leq y_2]$$

- Result 5.1. If Y_1 and Y_2 are independent random variables, then for any two constants a_1 and a_2 the random variable

$$W = a_1 Y_1 + a_2 Y_2$$

has mean and variance

$$E(W) = a_1 E(Y_1) + a_2 E(Y_2)$$

$$\text{Var}(W) = a_1^2 \text{Var}(Y_1) + a_2^2 \text{Var}(Y_2)$$

Fundamentals

- Result 5.2. If Y_1 and Y_2 are independent random variables that are **normally** distributed, then

$$W = a_1 Y_1 + a_2 Y_2$$

is normally distributed with mean and variance given by
Result 5.1

- Corollary: If \bar{Y}_1 and \bar{Y}_2 are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , then

$$\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Fundamentals

- Result 5.3. If \bar{Y}_1 and \bar{Y}_2 are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and the same variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where

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

- Note: If $n_1 = n_2$ then

$$s_p^2 = (s_1^2 + s_2^2)/2$$

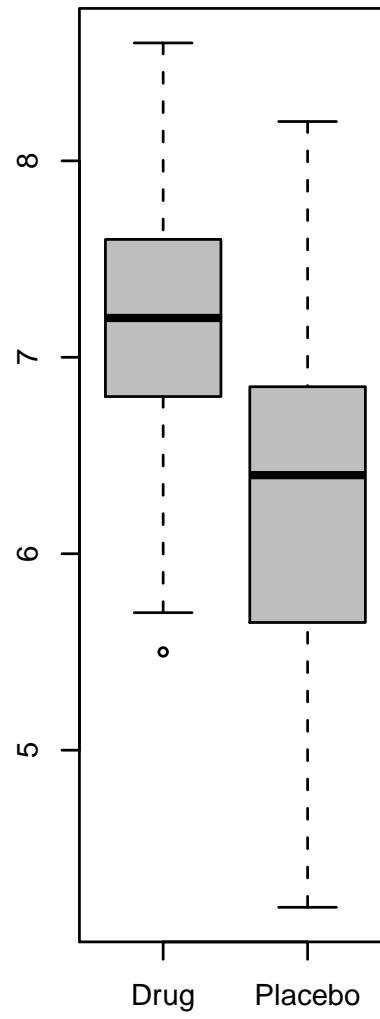
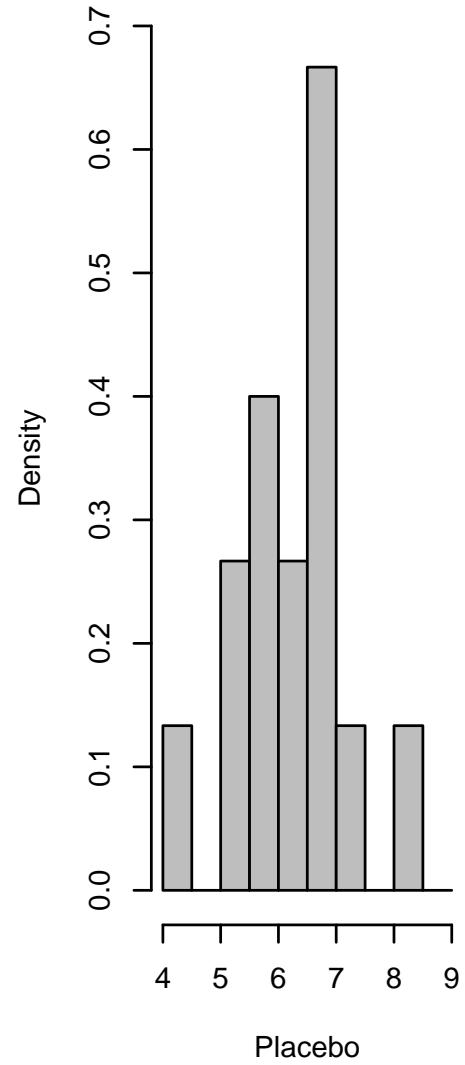
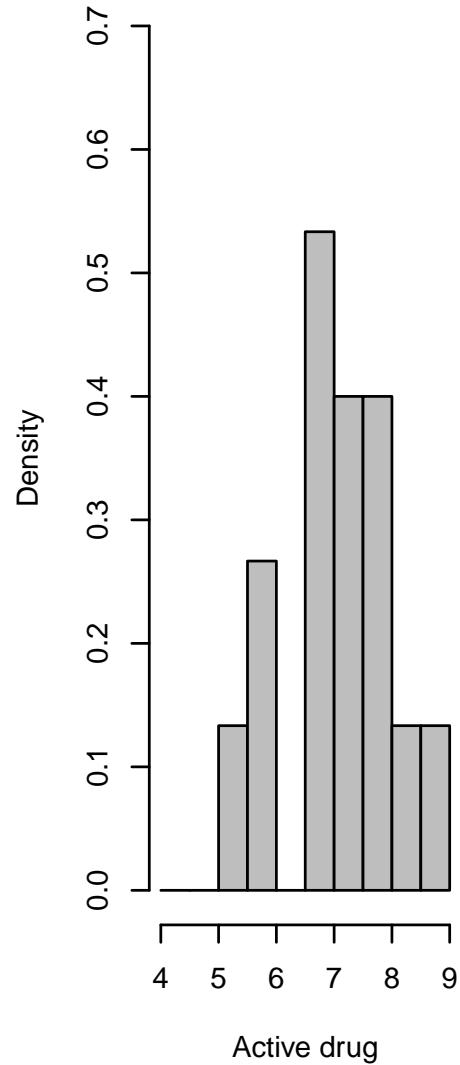
Two Sample t-test Example

- An experiment was conducted to see if a drug could prevent premature birth or low birthweight
- 30 women at risk of premature birth were randomly assigned to take the drug or a matching placebo (15 in each group)
- Endpoint: birthweight (in pounds)
- Let 1 = drug, 2 = placebo
- Consider $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 > \mu_2$
- $C_\alpha = \{t : t > t_{1-\alpha; 28}\}$
- $C_{0.05} = \{t : t > 1.7\}$

Two Sample t-test Example cont.

Drug	Placebo
6.9	6.4
7.6	6.7
7.3	5.4
7.6	8.2
6.8	5.3
7.2	6.6
8.0	5.8
5.5	5.7
5.8	6.2
7.3	7.1
8.2	7.0
6.9	6.9
6.8	5.6
5.7	4.2
8.6	6.8

Two Sample t-test Example cont.



Two Sample t-test Example cont.

- $\bar{y}_1 = 7.08, s_1 = 0.899$

- $\bar{y}_2 = 6.26, s_2 = 0.961$

- Thus

$$s_p^2 = \frac{14(0.899)^2 + 14(0.961)^2}{28} = 0.8657$$

$$t = \frac{7.08 - 6.26}{0.930\sqrt{2/15}} = 2.41$$

- Because $t \in C_{0.05}$, reject H_0

$$p = 1 - F_{t_{28}}(2.41) = 0.011$$

Two Sample t-test Example cont.

- R

```
> t.test(bw$drug,bw$placebo,var.equal=TRUE,alternative="greater")
```

Two Sample t-test

data: bw\$drug and bw\$placebo

t = 2.4136, df = 28, p-value = 0.01129

alternative hypothesis: true difference in means is greater than 0

Two Sample t-test Example cont.

- SAS

```
proc ttest; class trt; var bw;
```

The TTEST Procedure

Variable: bw

trt	N	Mean	Std Dev	Std Err
drug	15	7.0800	0.8994	0.2322
placebo	15	6.2600	0.9605	0.2480
Diff (1-2)		0.8200	0.9304	0.3397

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	28	2.41	0.0226
Satterthwaite	Unequal	27.88	2.41	0.0226

Homogeneity of Variance

- Want to test

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_A : \sigma_1^2 \neq \sigma_2^2$$

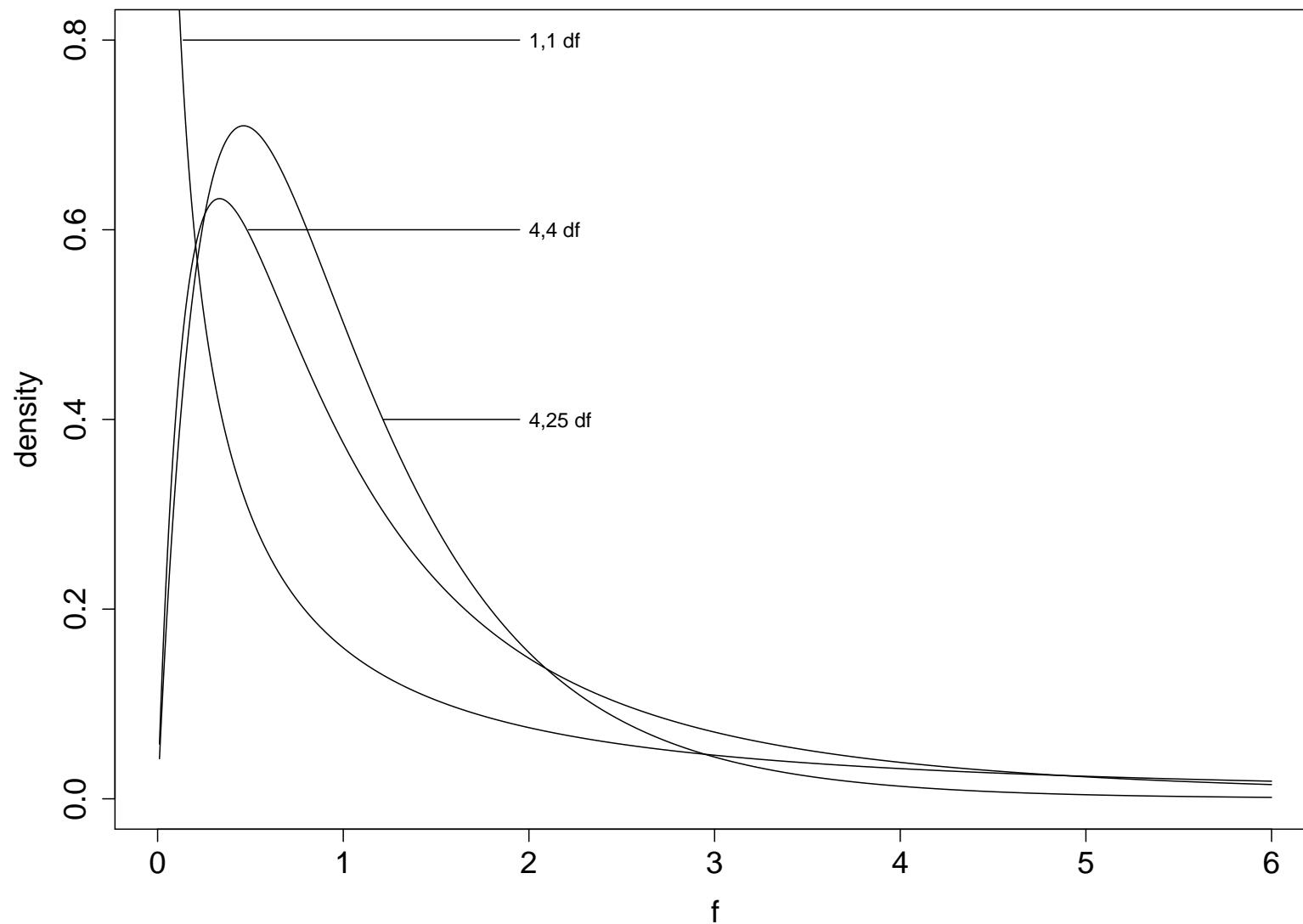
- We know that (assuming normality)

$$\frac{(n_k - 1)s_k^2}{\sigma_k^2} \sim \chi_{n_k - 1}^2 \text{ for } k = 1, 2$$

- If X_1 and X_2 are independent random variables with $X_1 \sim \chi_{v_1}^2$ and $X_2 \sim \chi_{v_2}^2$, then

$$\frac{X_1/v_1}{X_2/v_2} \sim F_{v_1, v_2}$$

F Distribution



Homogeneity of Variance

- Let

$$X_k = \frac{(n_k - 1)s_k^2}{\sigma_k^2} \quad \text{for } k = 1, 2$$

- It follows that

$$Y = \frac{X_1/(n_1 - 1)}{X_2/(n_2 - 1)} \sim F_{n_1-1, n_2-1}$$

- Thus

$$Y = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}$$

Homogeneity of Variance

- Under $H_0 : \sigma_1^2 = \sigma_2^2$,

$$\frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

- For $H_A : \sigma_1^2 \neq \sigma_2^2$, reject null if s_1^2/s_2^2 is very large or very small (i.e., near zero)
- Formally,

$$C_\alpha = \{f : f < F_{n_1-1, n_2-1, \alpha/2} \text{ or } f > F_{n_1-1, n_2-1, 1-\alpha/2}\}$$

where $f = s_1^2/s_2^2$

Homogeneity of Variance

- Note: $F_{v_1, v_2, \alpha} = 1/F_{v_2, v_1, 1-\alpha}$
- Table A.5 and A.6 of the text for two-sided $\alpha = 0.10$ and $\alpha = 0.02$; see errata
- R

```
> qf(0.975, 14, 14)
[1] 2.978588
> qf(0.025, 14, 14)
[1] 0.3357296
> 1/qf(0.025, 14, 14)
[1] 2.978588
```

- SAS

```
data; y = finv(0.975, 14, 14);
```

Homogeneity of Variance: BW Example

- $H_0 : \sigma_1^2 = \sigma_2^2; H_A : \sigma_1^2 \neq \sigma_2^2$

- For $\alpha = 0.05$,

$$C_{0.05} = \{f : f < F_{14,14,0.025} \text{ or } f > F_{14,14,0.975}\}$$

$$= \{f : f < 0.34 \text{ or } f > 2.98\}$$

- Observed test statistic

$$f = \frac{0.8994^2}{0.9605^2} = 0.8768$$

- Therefore, do not reject H_0

$$p = 2 \times F_{14,14}(0.8768) = 0.809$$

Homogeneity of Variance: BW Example cont.

- SAS

```
proc ttest; class trt; var bw;
```

The TTEST Procedure

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	28	2.41	0.0226
Satterthwaite	Unequal	27.88	2.41	0.0226

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	14	14	1.14	0.8090

Homogeneity of Variance: BW Example cont.

- R

```
> var.test(bw$drug,bw$placebo)
```

F test to compare two variances

```
data: bw$drug and bw$placebo  
F = 0.8767, num df = 14, denom df = 14, p-value = 0.809  
alternative hypothesis: true ratio of variances is not equal to 1
```

Testing Homogeneity of Variance

- Cf. page 133 of text
- Genuine interest in whether variances equal
- With respect to testing $H_0 : \mu_1 = \mu_2$
 - For small samples, potential for type II error
 - For large samples, CLT/Slutsky
 - Adjustment for sequential testing
- For additional reading, see Moser and Stevens (*The American Statistician* 1992)

Effect on Testing $\mu_1 = \mu_2$

- What if $\sigma_1^2 \neq \sigma_2^2$ and unknown?
- Solutions
 1. Large sample approximation
 2. Normality: Welch-Satterthwaite approximation
(Behrens-Fisher problem)
 3. Transformation
 4. Nonparametric methods: Wilcoxon rank sum

Large Sample Approximation

- If n_1 and n_2 are large, homogeneity of variance assumption is not important
- Recall CLT plus Slutsky implies

$$\bar{Y} \stackrel{\sim}{\sim} N \left(\mu, \frac{s^2}{n} \right)$$

- Thus

$$\bar{Y}_1 - \bar{Y}_2 \stackrel{\sim}{\sim} N \left(\mu_1 - \mu_2, \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)$$

Large Sample Approximation

- Therefore, to test $H_0 : \mu_1 - \mu_2 = \delta$, we can use

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Under H_0 , $Z \stackrel{\text{d}}{\sim} N(0, 1)$
- Approximation gets better as $n_1, n_2 \rightarrow \infty$
- Generally, require $n_j \geq 25$ for $j = 1, 2$
- Note that the assumption that the Y s are normally distributed is no longer needed either (CLT)

Large Sample Approximation: Example

- A study was done to compare the percent body fat of third graders at schools on two Native American reservations: Tohona and Apache
- $H_0 : \mu_T = \mu_A$ vs. $H_A : \mu_T \neq \mu_A$
- $n_T = 63, n_A = 35$
- $C_{0.05} = \{z : |z| > 1.96\}$
- $\bar{y}_T = 37.9\%; s_T = 8.66; \bar{y}_A = 32.8\%; s_A = 6.88;$
$$z = \frac{37.9 - 32.8}{\sqrt{\frac{8.66^2}{63} + \frac{6.88^2}{35}}} = 3.2;$$
p=0.0014

Welch-Satterthwaite Approximation

- Assume normality; n_1, n_2 small; $\sigma_1^2 \neq \sigma_2^2$
- Statistic

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{\text{df}}$$

- Note 5.2 of text:

$$\text{df}_{\text{text}} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}} - 2$$

Welch-Satterthwaite Approximation

- Welch (Biometrika 1938), SAS and R

$$df_{\text{welch}} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

- Use $\lfloor df \rfloor$ if using tables

Welch-Satterthwaite Approximation: Example

- Premature birth example

$$n_1 = n_2 = 15$$

$$s_1 = 0.8994, s_2 = 0.9605$$

$$df_{text} = 29.86$$

$$df_{welch} = 27.88$$

Welch-Satterthwaite Approximation: R

- R

```
> t.test(bw$drug,bw$placebo,var.equal=FALSE,alternative="greater")
```

Welch Two Sample t-test

```
data: bw$drug and bw$placebo  
t = 2.4136, df = 27.88, p-value = 0.01131
```

alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:

0.2419592 Inf

sample estimates:

mean of x mean of y

7.08 6.26

Summary

Normal	Var known	Var equal	N large	Test statistic
✓	✓			$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$
✓		✓	✓	$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$
			✓	$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$
✓				$\frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{df}$
				Transform, nonparametrics

Outline

- Already done: parametric/large sample
- Wilcoxon rank sum test
 - Hodges-Lehmann estimator, CIs
- Permutation test
- Kolmogorov-Smirnov test

Wilcoxon (Mann-Whitney) Rank Sum Test

1. Assume $Y_{1j}, \dots, Y_{n_j j}$ iid $F_j(y)$; $j = 1, 2$

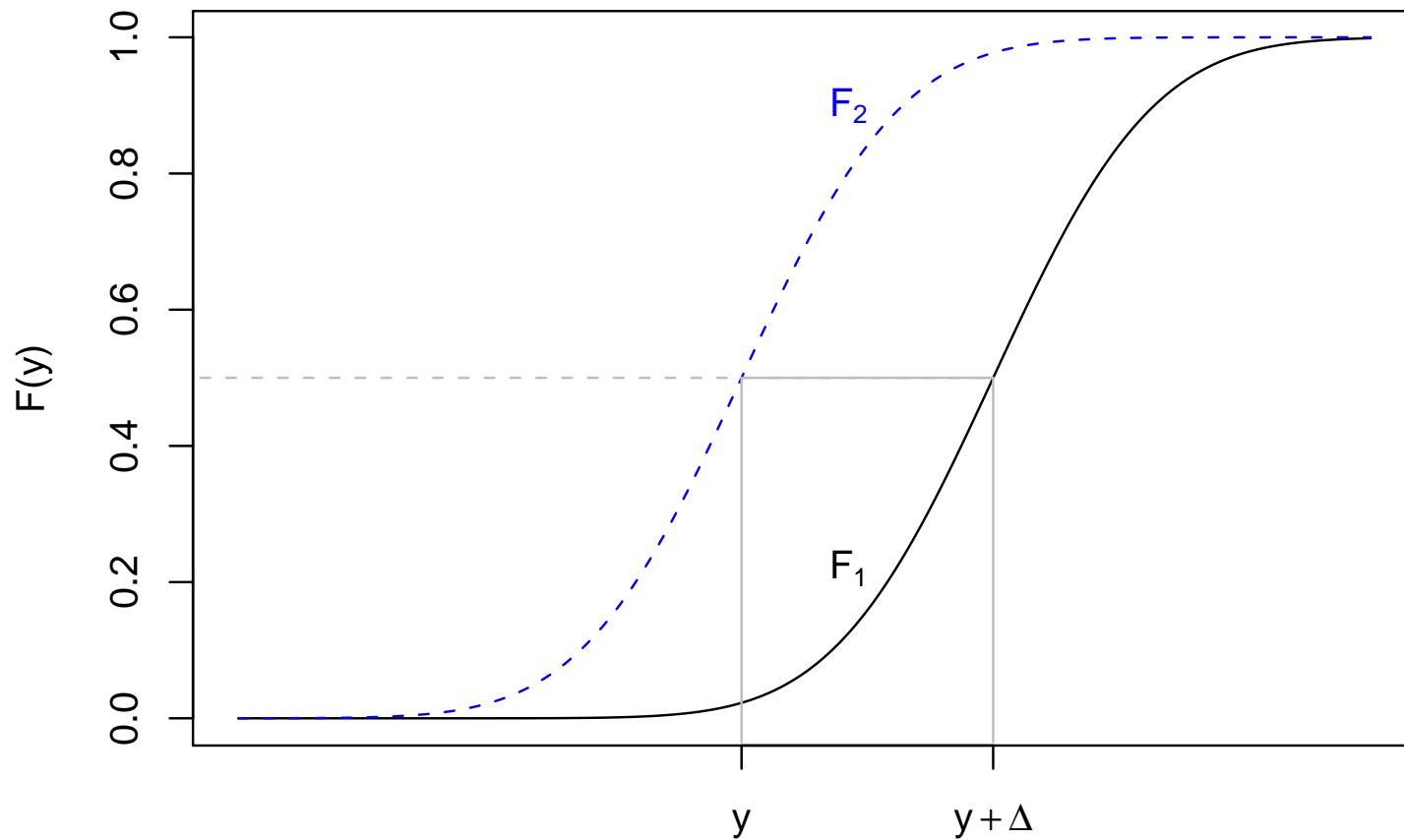
$$H_0 : F_1(y) = F_2(y)$$

$$H_A : F_1(y + \Delta) = F_2(y)$$

where Δ is a non-zero constant

2. Pool the two samples
3. Rank them from smallest to largest
4. Compute the sum of the ranks, W_1 , in group 1

Wilcoxon (Mann-Whitney) Rank Sum Test



Wilcoxon Rank Sum Test

- There are $N = n_1 + n_2$ subjects in our study
- Thus there are $\binom{N}{n_1}$ possible outcomes (in terms of which ranks are in group 1)
- Under H_0 , each is equally likely
- We compute the distribution of W_1 by enumeration

Wilcoxon Rank Sum Test: Example

- A new drug is being tested in humans for the first time to assess effect on CD4⁺ T cells in patients with HIV
- 7 individuals are randomized to 2 groups:

control ($n_1 = 3$) or drug ($n_2 = 4$)

- Endpoint is percent change in CD4⁺ count from baseline
- Null hypothesis is that the drug has no effect

$$H_0 : \Delta = 0$$

$$H_A : \Delta \neq 0$$

Wilcoxon Rank Sum Test: Example cont.

- Data: control (65, 73, 69); drug (89, 70, 92, 88)
- There are $\binom{7}{3}$ possible outcomes of the study
i.e., there are 35 possible sets of rankings for group 1

Wilcoxon Rank Sum Test: $n_1 = 3, n_2 = 4$

Ranks	W_1	Ranks	W_1	Ranks	W_1
1,2,3	6	1,5,6	12	2,6,7	15
1,2,4	7	1,5,7	13	3,4,5	12
1,2,5	8	1,6,7	14	3,4,6	13
1,2,6	9	2,3,4	9	3,4,7	14
1,2,7	10	2,3,5	10	3,5,6	14
1,3,4	8	2,3,6	11	3,5,7	15
1,3,5	9	2,3,7	12	3,6,7	16
1,3,6	10	2,4,5	11	4,5,6	15
1,3,7	11	2,4,6	12	4,5,7	16
1,4,5	10	2,4,7	13	4,6,7	17
1,4,6	11	2,5,6	13	5,6,7	18
1,4,7	12	2,5,7	14		

Wilcoxon Rank Sum Test: $n_1 = 3, n_2 = 4$

w	$\Pr[W = w]$	$F(w)$	$\Pr[W \geq w]$
6	0.0286	0.0286	1
7	0.0286	0.0571	0.9714
8	0.0571	0.1143	0.9429
9	0.0857	0.2000	0.8857
10	0.1143	0.3143	0.8000
11	0.1143	0.4286	0.6857
12	0.1429	0.5714	0.5714
13	0.1143	0.6857	0.4286
14	0.1143	0.8000	0.3143
15	0.0857	0.8857	0.2000
16	0.0571	0.9429	0.1143
17	0.0286	0.9714	0.0571
18	0.0286	1	0.0286

Wilcoxon Rank Sum Test: $n_1 = 3, n_2 = 4$

- Note that it is impossible to reject H_0 for a two-sided alternative when $\alpha = 0.05$
- For a two-sided $\alpha = 0.1$ test

$$C_\alpha = \{6, 18\}$$

- Observed $W_1 = 1 + 2 + 4 = 7$
- So do not reject H_0

Wilcoxon Rank Sum Test

- Note

$$W_1 + W_2 = \sum_{i=1}^N i = \frac{N(N+1)}{2}$$

- Thus, under H_0

$$E(W_1) = \frac{n_1}{N} \frac{N(N+1)}{2} = \frac{n_1(N+1)}{2}$$

- Similarly

$$\text{Var}(W_1) = \frac{n_1 n_2 (N+1)}{12}$$

(cf. Lehmann 1998, Example 3, p 332)

Large Sample Approximation

- If n_1 and n_2 are large

$$Z = \frac{W_1 - E(W_1)}{\sqrt{\text{Var}(W_1)}}$$

will be approximately $N(0, 1)$

- Approximation is good for $n_1, n_2 \geq 12$
- If there are ties

$$\text{Var}(W_1) = \frac{n_1 n_2 (N + 1)}{12} - \frac{n_1 n_2}{12 N (N - 1)} \sum_{i=1}^q t_i (t_i - 1) (t_i + 1)$$

Wilcoxon Rank Sum Test: BW Example

Drug	Rank	Placebo	Rank
5.5	4	4.2	1
5.7	6.5	5.3	2
5.8	8.5	5.4	3
6.8	15	5.6	5
6.8	15	5.7	6.5
6.9	18	5.8	8.5
6.9	18	6.2	10
7.2	22	6.4	11
7.3	23.5	6.6	12
7.3	23.5	6.7	13
7.6	25.5	6.8	15
7.6	25.5	6.9	18
8.0	27	7.0	20
8.2	28.5	7.1	21
8.6	30	8.2	28.5

Wilcoxon Rank Sum Test: BW Example cont.

- $H_0 : \Delta = 0; \quad H_A : \Delta > 0$
- $C_{0.05} = \{z : z > 1.645\}$
- $E(W_1) = \frac{15(31)}{2} = 232.5$
- $\text{Var}(W_1 | \text{no ties}) = \frac{15^2(31)}{12} = 581.25$

Wilcoxon Rank Sum Test: BW Example cont.

- Tie correction:

$$q = 7; \quad t_1 = t_2 = 2; \quad t_3 = t_4 = 3; \quad t_5 = t_6 = t_7 = 2$$

$$\sum_{i=1}^q t_i(t_i - 1)(t_i + 1) = 78$$

$$\text{Var}(W_1) = 581.25 - \frac{78(15)^2}{12(30)(29)} = 579.57$$

Wilcoxon Rank Sum Test: BW Example cont.

- $w_1 = 290.5$

$$z = \frac{290.5 - 232.5}{\sqrt{579.57}} = 2.409;$$

- Reject H_0
- $p = 1 - \Phi(2.409) = 0.008$
- Note: without tie correction $z = 2.406$; $p = 0.008$

Wilcoxon Rank Sum Test: BW Example cont.

- SAS

```
proc npar1way wilcoxon correct=no; class trt; var bw;
```

Wilcoxon Scores (Rank Sums) for Variable bw
Classified by Variable trt

trt	N	Sum of	Expected	Std Dev	Mean
		Scores	Under H0	Under H0	Score
<hr/>					
drug	15	290.50	232.50	24.074239	19.366667
plac	15	174.50	232.50	24.074239	11.633333

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic (S)	290.5000
Normal Approximation	
Z	2.4092
One-Sided Pr > Z	0.0080
Two-Sided Pr > Z	0.0160

Wilcoxon Rank Sum Test: BW Example cont.

- R

```
> wilcox.test(bw$drug,bw$placebo,alternative="greater",exact=F,correct=F)
```

```
Wilcoxon rank sum test
```

```
data: bw$drug and bw$placebo
W = 170.5, p-value = 0.007993
alternative hypothesis: true mu is greater than 0
```

- Note that `w` here is not the sum of the ranks; the slides on the Mann-Whitney test will explain what it is

Wilcoxon Rank Sum Exact P-values

- For a two-sided alternative, exact p-values are computed (under the null) by

$$\Pr [|W_1 - E(W_1)| \geq |w_1 - E(W_1)|]$$

where

$$E(W_1) = \frac{n_1(N + 1)}{2}$$

- Without ties, the distribution of W_1 is symmetric about $E(W_1)$

Wilcoxon Rank Sum Exact P-values: Example

- Suppose $\mathbf{Y}_1 = (65, 70, 73)$ and $\mathbf{Y}_2 = (70, 89)$
- There are $\binom{5}{2} = 10$ possible rankings for group 1

Ranks	W_1	$ W_1 - E(W_1) $	Ranks	W_1	$ W_1 - E(W_1) $
1,2,5,2,5	6	3	1,4,5	10	1
1,2,5,4	7.5	1.5	2,5,2,5,4	9	0
1,2,5,4	7.5	1.5	2,5,2,5,5	10	1
1,2,5,5	8.5	0.5	2,5,4,5	11.5	2.5
1,2,5,5	8.5	0.5	2,5,4,5	11.5	2.5

- Thus $|w_1 - E(W_1)| = |7.5 - 9| = 1.5$, giving $p = 0.5$

Wilcoxon Rank Sum Exact P-values: R

```
> # First need to install package exactRankTests  
> wilcox.exact(c(65,70,73),c(70,89))
```

```
Exact Wilcoxon rank sum test  
  
data: c(65, 70, 73) and c(70, 89)  
W = 1.5, p-value = 0.5  
alternative hypothesis: true mu is not equal to 0
```

Wilcoxon Rank Sum Exact P-values: SAS

```
proc npar1way wilcoxon; class group; var y;  
    exact wilcoxon;
```

Wilcoxon Scores (Rank Sums) for Variable y
Classified by Variable group

group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
<hr/>					
1	3	7.50	9.0	1.688194	2.500
2	2	7.50	6.0	1.688194	3.750

Average scores were used for ties.

Wilcoxon Two-Sample Test

Statistic (S) 7.5000

Exact Test

One-Sided Pr >= S 0.3000

Two-Sided Pr >= |S - Mean| 0.5000

Mann-Whitney Test

- Consider all $n_1 n_2$ possible pairs

$$(Y_{1i}, Y_{2j}); i = 1, 2, \dots, n_1; j = 1, 2, \dots, n_2$$

- Let U_1 equal the number of pairs with $Y_{1i} < Y_{2j}$
- It can be shown that

$$U_1 = \frac{n_1(N + n_2 + 1)}{2} - W_1$$

- Reject $H_0 : \Delta = 0$ in favor of $H_A : \Delta > 0$ if W_1 large, i.e., U_1 small

Mann-Whitney Test

- That is, the Mann-Whitney and Wilcoxon rank sum test are equivalent
- This explains the R output

$$U_2 = \frac{n_2(N + n_1 + 1)}{2} - W_2 = \frac{15 \times 46}{2} - 174.5 = 170.5$$

- Reject $H_0 : \Delta = 0$ in favor of $H_A : \Delta > 0$ if W_2 small, i.e., U_2 large

Mann-Whitney Test

- Table A.10 in the text gives critical values

For example, $n_1 = n_2 = 15$ and a one-sided test

$\alpha = 0.01$: critical value = 169

$\alpha = 0.005$: critical value = 174

- Efficiency

Compared to the t-test, the ARE = 0.955 under normality;

never worse than 0.864

Hodges-Lehmann Estimator

- Assume $F_1(y + \Delta) = F_2(y)$ for some constant Δ .
- For the Wilcoxon rank sum test:

$$H_0 : \Delta = 0$$

$$H_A : \Delta \neq 0$$

- If we reject H_0 , we may want an estimate of Δ
- Estimate Δ by the amount $\hat{\Delta}$ by which the Y_{2j} s must be shifted to give the best possible agreement with the Y_{1i} s

Hodges-Lehmann Estimator

- From the Mann-Whitney perspective, we want

$$Y_{1i} > Y_{2j} + \hat{\Delta} \text{ half of the time}$$

- Thus

$$\hat{\Delta} = \text{median}\{Y_{1i} - Y_{2j} : i = 1, \dots, n_1; j = 1, \dots, n_2\}$$

- CI for Δ ?

CIs by Inverting a Test

- For each possible value of $\theta_0 \in \Omega$, let $C_\alpha(\theta_0)$ denote the critical region for testing $H_0 : \theta = \theta_0$ at the α level of significance
- Let X denote the corresponding test statistic and set

$$S(X) = \{\theta : X \notin C_\alpha(\theta)\}$$

- Claim: $S(X)$ is a $(1 - \alpha) \times 100\%$ CI for θ
- Proof:

$$\Pr_\theta[\theta \in S(X)] = \Pr_\theta[X \notin C_\alpha(\theta)] \geq 1 - \alpha$$

CIs by Inverting a Test

- For given value x of a test statistic X , find all values of θ for which we would not reject H_0 at the α level of significance
- For example, consider the Wilcoxon Rank Sum test:
For a given value of W_1 , find all values of Δ for which we would fail to reject H_0

Rank Sum Test: BW Example in R

```
> wilcox.test(bw$drug,bw$placebo,exact=F,correct=F,conf.int=T)
```

```
Wilcoxon rank sum test

data: bw$drug and bw$placebo
W = 170.5, p-value = 0.01599
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.1000055 1.5000265
sample estimates:
difference in location
0.8000382
```

```
> median(outer(bw$drug,bw$placebo,"-"))
[1] 0.8
```

Rank Sum Test: BW Example in R cont.

```
> # Obtain CI by inverting test  
> # Note that the original observations are to 1 decimal place
```

```
> wilcox.test(bw$drug,bw$placebo+.1,exact=F,correct=F)
```

W = 165, p-value = 0.02927

```
> wilcox.test(bw$drug,bw$placebo+.10001,exact=F,correct=F)
```

W = 159, p-value = 0.05366

```
> wilcox.test(bw$drug,bw$placebo+1.5,exact=F,correct=F)
```

W = 68, p-value = 0.06442

```
> wilcox.test(bw$drug,bw$placebo+1.50001,exact=F,correct=F)
```

W = 63, p-value = 0.03997

Permutation Test

- Cf. section 8.9 of the text
- $H_0 : F_1 = F_2$
- Test statistic $D \equiv \bar{Y}_1 - \bar{Y}_2$
- Suppose N subjects randomly assigned to two groups of sizes n_1, n_2
- There are $\binom{N}{n_1}$ possible group assignments of sizes n_1, n_2 and each is equally likely under H_0
- Each of these assignments results in a value of $\bar{Y}_1 - \bar{Y}_2$
- Compute $\bar{Y}_1 - \bar{Y}_2$ for each possible assignment

Permutation Test

- Compute the CDF of $\bar{Y}_1 - \bar{Y}_2$ under $H_0 : F_1 = F_2$
- From the CDF, determine the critical region
- Example: HIV study $\binom{7}{3} = 35$ possible group assignments into groups of sizes 3 and 4

Example: All Possible Group Assignments

Group 1	Group 2	$\bar{Y}_1 - \bar{Y}_2$	Group 1	Group 2	$\bar{Y}_1 - \bar{Y}_2$
65 69 70	73 88 89 92	-17.50	65 69 73	70 88 89 92	-15.75
65 69 88	70 73 89 92	-7.00	65 69 89	70 73 88 92	-6.42
65 69 92	70 73 88 89	-4.67	65 70 73	69 88 89 92	-15.17
65 70 88	69 73 89 92	-6.42	65 70 89	69 73 88 92	-5.83
65 70 92	69 73 88 89	-4.08	65 73 88	69 70 89 92	-4.67
65 73 89	69 70 88 92	-4.08	65 73 92	69 70 88 89	-2.33
65 88 89	69 70 73 92	4.67	65 88 92	69 70 73 89	6.42
65 89 92	69 70 73 88	7.00	69 70 73	65 88 89 92	-12.83
69 70 88	65 73 89 92	-4.08	69 70 89	65 73 88 92	-3.50
69 70 92	65 73 88 89	-1.75	69 73 88	65 70 89 92	-2.33
69 73 89	65 70 88 92	-1.75	69 73 92	65 70 88 89	0.00
69 88 89	65 70 73 92	7.00	69 88 92	65 70 73 89	8.75
69 89 92	65 70 73 88	9.33	70 73 88	65 69 89 92	-1.75
70 73 89	65 69 88 92	-1.17	70 73 92	65 69 88 89	0.58
70 88 89	65 69 73 92	7.58	70 88 92	65 69 73 89	9.33
70 89 92	65 69 73 88	9.92	73 88 89	65 69 70 92	9.33
73 88 92	65 69 70 89	11.08	73 89 92	65 69 70 92	10.67
88 89 92	65 69 70 73	20.42			

EDF of $\bar{Y}_1 - \bar{Y}_2$

d	$\Pr[\bar{Y}_1 - \bar{Y}_2 \leq d]$	$\Pr[\bar{Y}_1 - \bar{Y}_2 \geq d]$	d	$\Pr[\bar{Y}_1 - \bar{Y}_2 \leq d]$	$\Pr[\bar{Y}_1 - \bar{Y}_2 \geq d]$
-17.50	0.029	1.000	6.41	0.686	0.343
-15.75	0.057	0.971	7.00	0.743	0.314
-15.17	0.086	0.943	7.58	0.771	0.257
-12.83	0.114	0.914	8.75	0.800	0.229
-7.00	0.143	0.886	9.33	0.886	0.200
-6.42	0.200	0.857	9.92	0.914	0.114
-5.83	0.229	0.800	10.67	0.943	0.086
-4.67	0.286	0.771	11.08	0.971	0.057
-4.08	0.371	0.714	20.42	1.000	0.029
-3.50	0.400	0.629			
-2.33	0.457	0.600			
-1.75	0.543	0.543			
-1.17	0.571	0.457			
0.00	0.600	0.429			
0.58	0.629	0.400			
4.67	0.657	0.371			

Permutation Test Example

- Symmetric critical region for $\alpha = 0.1$

$$C_{0.10} = \{D : D = -17.5 \text{ or } D = 20.42\}$$

where $D = \bar{Y}_1 - \bar{Y}_2$

- Observed $d = -15.75$
- So do not reject H_0

Permutation Test

- No assumptions except random assignment
- Computations are extensive if N is moderately large
e.g., for $N = 20$, the number of permutations is $20!$,
which is $> 2 \times 10^{18}$
However, $\binom{20}{10} = 184,756$
- *Conditional test* (conditional on the observed Y s), that is, Y s fixed
- Exact: probability of rejecting the null when it holds never exceeds the nominal significance level

Kolmogorov-Smirnov Test

- Want to test

$$H_0 : F_1(y) = F_2(y) \text{ for all } y$$

versus general alternative

$$H_A : F_1(y) \neq F_2(y) \text{ for at least one } y$$

- KS test

$$D = \max_y |F_{1n}(y) - F_{2m}(y)|$$

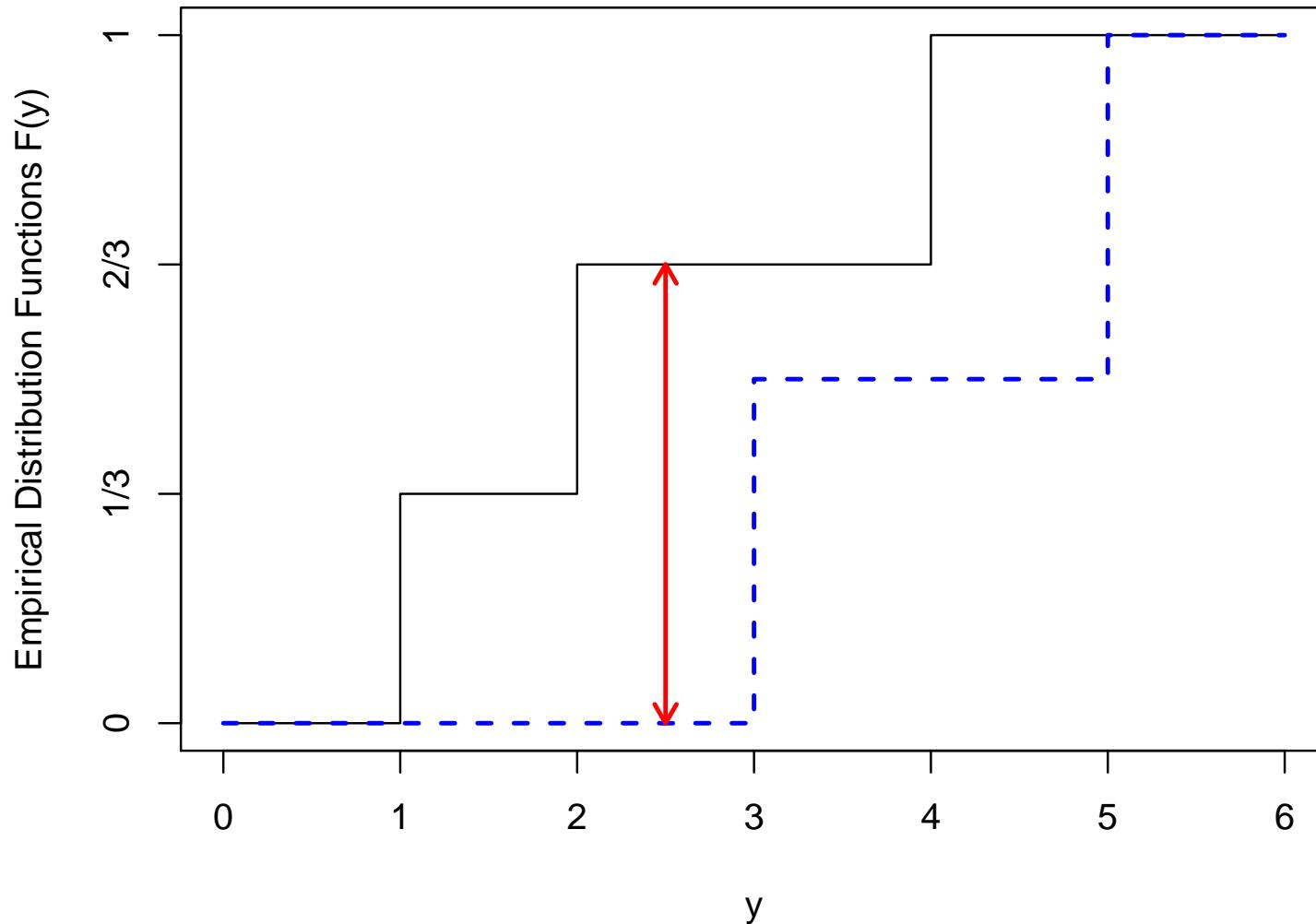
where $F_{1n}(y)$ and $F_{2m}(y)$ are the EDFs for the samples from F_1 and F_2 , respectively

- Note that even though the alternative is two-sided, we reject H_0 only for large values of D

Kolmogorov-Smirnov Test

- Can be viewed as a rank test (text p. 279)
- Large sample based critical values on p. 268 of the text
- For small samples, exact distribution based on enumeration of all possible group assignments as illustrated in the following example
- Example: $\mathbf{Y}_1 = (1, 2, 4), \mathbf{Y}_2 = (3, 5)$

Kolmogorov-Smirnov Test: Example



Kolmogorov-Smirnov Test: Example cont.

- Observe $d = 2/3$
- There are 10 possible group assignments

\mathbf{Y}_1	D	\mathbf{Y}_1	D
1,2,3	1	1,4,5	2/3
1,2,4	2/3	2,3,4	1/2
1,2,5	2/3	2,3,5	1/2
1,3,4	1/2	2,4,5	2/3
1,3,5	1/3	3,4,5	1

- Thus $p = \Pr[D \geq d] = 0.6$

KS in R and SAS

- R

```
> ks.test(c(1,2,4),c(3,5))
```

```
Two-sample Kolmogorov-Smirnov test
```

```
data: c(1, 2, 4) and c(3, 5)
D = 0.6667, p-value = 0.6
alternative hypothesis: two-sided
```

- SAS

```
proc npar1way;
  class group;
  var x;
  exact ks;
run;
```

KS in SAS

Kolmogorov-Smirnov Test for Variable x
Classified by Variable group

group	N	EDF at	Deviation from Mean
		Maximum	at Maximum
1	3	0.666667	0.461880
2	2	0.000000	-0.565685
Total	5	0.400000	

Maximum Deviation Occurred at Observation 2

Value of x at Maximum = 2.0

Kolmogorov-Smirnov Two-Sample Test

D = max |F1 - F2| 0.6667

Asymptotic Pr > D 0.6604

Exact Pr >= D 0.6000

D+ = max (F1 - F2) 0.6667

Asymptotic Pr > D+ 0.3442

Exact Pr >= D+ 0.3000

D- = max (F2 - F1) 0.0000

Asymptotic Pr > D- 1.0000

Exact Pr >= D- 1.0000

Discussion

- Wilcoxon rank sum test: Default non-parametric test
- Permutation test
 - Asymptotically equivalent to t-test; thus most powerful asymptotically under normality
 - Computationally intensive because unique to each data set
 - Sensitive to outliers
- Kolmogorov-Smirnov test
 - Employ if trying to detect difference in distributions other than location shift

BIOS 662 Fall 2018

Count Data

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- One sample binary outcome
- Two sample binary outcome
- Measures of association
- Confounding - Mantel-Haenszel
- Matching - McNemar

Binomial Random Variable

- $X_1, \dots, X_n \sim \text{Bernoulli}(\pi)$
- $Y = \sum_{i=1}^n X_i \sim \text{Binomial}(n, \pi)$
- Four key conditions
 1. Binary response (0/1)
 2. Observed a known number of times n
 3. Success probability (π) the same each time
 4. Independence between trials
- Example 6.1 in the text: Smoke exposure

Binomial Random Variable

- Hypothesis testing

$$H_0 : \pi = \pi_0 \text{ vs. } H_A : \pi \neq \pi_0$$

- The statistic Y is the count of the successes
- Under the null, $Y \sim \text{Binomial}(n, \pi_0)$
- Need to find $y_{\alpha/2}$ and $y_{1-\alpha/2}$ such that

$$\Pr[Y \leq y_{\alpha/2} | H_0] \leq \alpha/2$$

and

$$\Pr[Y \geq y_{1-\alpha/2} | H_0] \leq \alpha/2$$

Exact Test for Binomial Proportion

- For small samples, compute exact CR using

$$\Pr[Y \leq y_{\alpha/2}] = \sum_{i=0}^{y_{\alpha/2}} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$$

$$\Pr[Y \geq y_{1-\alpha/2}] = \sum_{i=y_{1-\alpha/2}}^n \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$$

- Binomial probabilities are computed or read from a table;
e.g., in R using `pbinom` or `dbinom`;
in SAS using `CDF('BINOMIAL', m, p, n)`
where `m` is the number of successes

Exact Test for Binomial: Example

- Suppose $n = 12$, $\pi_0 = 0.4$, $\alpha = 0.05$

y	$\Pr[Y \leq y]$	$\Pr[Y \geq y]$
0	0.00218	1.00000
1	0.01959	0.99782
2	0.08344	0.98041
:	:	
7	0.94269	0.15821
8	0.98473	0.05731
9	0.99719	0.01527
10	0.99968	0.00281
11	0.99998	0.00032
12	1.00000	0.00002

- Thus $y_{0.025} = 1$, $y_{0.975} = 9$, and

$$C_{0.05} = \{Y : Y \leq 1 \text{ or } Y \geq 9\}$$

Exact Test for Binomial: Example II

- Suppose it is known that the 1-year death rate for a particular form of cancer is 30%.
- A new therapy designed to decrease the death rate is to be tried on 15 patients

$$H_0 : \pi = 0.3 \quad \text{vs.} \quad H_A : \pi < 0.3$$

- Then want $C_\alpha = \{Y : Y \leq y_\alpha\}$ where

$$\sum_{i=0}^{y_\alpha} \binom{15}{i} 0.3^i 0.7^{15-i} \leq \alpha$$

- From table or R:

$$C_{0.05} = \{Y : Y \leq 1\} = \{Y : Y \in \{0, 1\}\}$$

Binomial: Large Sample

- Test of hypothesis for binomial data when n is large
- Normal approximation to binomial
- If $Y \sim \text{Binomial}(n, \pi)$, then for large n the distribution of

$$Z = \frac{Y - n\pi}{\sqrt{n\pi(1 - \pi)}}$$

is approximately $N(0, 1)$

- Approximation improves as $n \rightarrow \infty$
- Rule of thumb: $n\pi(1 - \pi) \geq 10$

Binomial: Example

- Revisit cancer example: Now suppose we test the new therapy on 150 patients
- Then

$$C_{0.05} = \{z : z < -1.645\}$$

where

$$Z = \frac{Y - 45}{\sqrt{150(0.3)(0.7)}}$$

Binomial: Small Sample CIs

- Invert the exact test: Find all π_0 such that $H_0 : \pi = \pi_0$ would not be rejected
- To get an exact $100(1 - \alpha)\%$ CI for π , solve these equations for π_L and π_U :

$$\Pr[Y \geq y | \pi = \pi_L] = \sum_{k=y}^n \binom{n}{k} \pi_L^k (1 - \pi_L)^{n-k} = \alpha/2$$

$$\Pr[Y \leq y | \pi = \pi_U] = \sum_{k=0}^y \binom{n}{k} \pi_U^k (1 - \pi_U)^{n-k} = \alpha/2$$

- Known as the *Clopper-Pearson* interval

Binomial: Small Sample CIs

- Can show that

$$\pi_L = \frac{y}{y + (n - y + 1) \times F_{2(n-y+1), 2y, 1-\alpha/2}}$$

for $1 \leq y \leq n$ ($\pi_L = 0$ for $y = 0$); and

$$\pi_U = \frac{y + 1}{y + 1 + (n - y)/F_{2(y+1), 2(n-y), 1-\alpha/2}}$$

for $0 \leq y \leq n - 1$ ($\pi_U = 1$ for $y = n$)

- This CI can be “extremely conservative”; cf. Wypij (*Encyclopedia of Biostatistics*, 1998)

Binomial: Small Sample CIs

- For example, suppose $n = 12$ and $y = 4$

- Then

$$\pi_L = \frac{4}{4 + 9 \times F_{18,8,0.975}}$$

- R

```
> 4/(4+9*qf(0.975,18,8))  
[1] 0.0992461
```

- SAS

```
data; x=4/(4+9*quantile('f',0.975,18,8));
```

Binomial: Small Sample CIs

- R code

```
> binom.test(4,12)

Exact binomial test

data: 4 and 12

number of successes = 4, number of trials = 12, p-value = 0.3877

alternative hypothesis: true probability of success is not equal to 0.5

95 percent confidence interval:

0.0992461 0.6511245
```

Binomial: Small Sample CIs

- SAS code

```
data; input event count; datalines;  
0 4  
1 8  
;  
  
proc freq; tables event; exact binomial; weight count; run;  
  
Binomial Proportion for event = 0  
-----  
Proportion (P)          0.3333  
ASE                   0.1361  
95% Lower Conf Limit  0.0666  
95% Upper Conf Limit  0.6001  
  
Exact Conf Limits  
95% Lower Conf Limit  0.0992  
95% Upper Conf Limit  0.6511
```

Binomial: Small Sample CIs

- Suppose $y = 0$
- Then $\pi_L = 0$ because

$$\Pr[Y \geq 0 | \pi = \pi_L] = \sum_{k=0}^n \binom{n}{k} \pi_L^k (1 - \pi_L)^{n-k} = 1$$

for any $\pi_L \neq 0$

- For the upper bound

$$\Pr[Y \leq 0 | \pi = \pi_U] = \sum_{k=0}^0 \binom{n}{k} \pi_U^k (1 - \pi_U)^{n-k} = \alpha/2$$

implies $\pi_U = 1 - (\alpha/2)^{1/n}$

Binomial: Small Sample CIs

- Suppose $n = 10$, $\alpha = 0.05$, $y = 0$
- $\pi_L = 0$, $\pi_U = 1 - 0.025^{1/10} = 0.3085$
- R

```
> binom.test(0,10)
```

```
Exact binomial test
```

```
data: 0 and 10
number of successes = 0, number of trials = 10, p-value = 0.001953
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
0.0000000 0.3084971
```

Binomial: Large Sample CIs

- Let $p = Y/n$ where Y is the number of successes in n trials
- Can think of this as a random sample X_1, X_2, \dots, X_n in which $X_i = 1$ for a success and 0 otherwise, with $Y = \sum_1^n X_i$, and so $p = \bar{X}$
- If n is sufficiently large,

$$p \sim N\left(\pi, \frac{\pi(1 - \pi)}{n}\right)$$

- Thus an approximate $100(1 - \alpha)\%$ CI for π is

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1 - p)}{n}}$$

- Rule of thumb: $np(1 - p) \geq 10$

Binomial: Example

- Suppose a random sample of 886 undergrads at a college finds that 321 report binge drinking at least once in the past year
- Then point estimate for π is

$$p = \frac{321}{886} = 0.36$$

- An approximate 95% CI for the proportion of binge drinkers is:

$$0.36 \pm 1.96 \sqrt{\frac{(0.36)(0.64)}{886}} = 0.36 \pm 0.03 = (0.33, 0.39)$$

Comparing Two Proportions

- Small sample sizes
 - Fisher's exact test
- Large sample sizes
 - normal approximation to the binomial
 - χ^2 test

Comparing Two Proportions

- Put the data in a 2×2 table

	Success	Failure	
Sample 1	n_{11}	n_{12}	n_1
Sample 2	n_{21}	n_{22}	n_2
	m_1	m_2	N

- Suppose $n_{11} \sim \text{Binomial}(n_1, \pi_1)$

and $n_{21} \sim \text{Binomial}(n_2, \pi_2)$

- Hypotheses

$$H_0 : \pi_1 = \pi_2$$

versus

$$H_A : \pi_1 \neq \pi_2 \quad \text{or} \quad H_A : \pi_1 < \pi_2$$

Fisher's Exact Test

- Assume the margins m_1, m_2, n_1, n_2 are fixed
- Then once we know n_{11} , the other values n_{12}, n_{21} , and n_{22} are uniquely determined
- Under H_0 , can show

$$\begin{aligned}\Pr[n_{11} = k | m_1, n_1, n_2] &= \frac{\binom{n_1}{k} \binom{n_2}{m_1 - k}}{\binom{N}{m_1}} \\ &= \frac{n_1! n_2! m_1! m_2!}{N! n_{11}! n_{12}! n_{21}! n_{22}!}\end{aligned}$$

- This is the *hypergeometric* distribution

Fisher's Exact Test

- For Fisher's exact test, we use the hypergeometric distribution
 1. Rearrange the table so that the row with the smaller row total is the first row and the column with the smaller column total is the first column
 2. Set $n_{11} = 0$ and compute $\Pr[n_{11} = 0]$ using the hypergeometric distribution
 3. Construct the next table by increasing n_{11} by 1 and re-compute the probability
 4. Repeat step 3 until one of the remaining 3 cells is 0
 5. This gives the CDF for n_{11}

Fisher's Exact Test: Example

- A study compared the surgical mortality for patients receiving an emergency coronary bypass with those receiving a non-emergency bypass

	Dead	Alive	
Emergency	1	19	20
Non-emergency	7	369	376
Total	8	388	396

- Null hypothesis

$$H_0 : \Pr[\text{dead} | \text{emergency}] = \Pr[\text{dead} | \text{non-emergency}]$$

$$H_0 : \pi_1 = \pi_2$$

Fisher's Exact Test: Example cont.

- Set $n_{11} = 0$

	Dead	Alive	
Emergency	0	20	20
Non-emergency	8	368	376
Total	8	388	396

$$\Pr[n_{11} = 0 \mid \text{observed margins}] = \frac{20! 376! 388! 8!}{396! 0! 20! 8! 368!} = 0.658$$

- Similarly for $\Pr[n_{11} = 1]$, $\Pr[n_{11} = 2]$, ...

Fisher's Exact Test: Example cont.

a	$\Pr[n_{11} = a]$	$\Pr[n_{11} \leq a]$	$\Pr[n_{11} \geq a]$
0	0.658	0.658	1.000
1	0.285	0.943	0.342
2	0.051	0.994	0.057
3	0.005	0.999	0.006
4	<0.001	>0.999	<0.001
5	<0.001	>0.999	<0.001
6	<0.001	>0.999	<0.001
7	<0.001	>0.999	<0.001
8	<0.001	1.000	<0.001

Fisher's Exact Test: Example cont.

- If $H_A : \pi_1 > \pi_2$, we would reject H_0 for large n_{11}
- For example

$$C_{0.05} = \{n_{11} : n_{11} \geq 3\}$$

- P-value for this study

$$\Pr[n_{11} \geq 1] = 1 - 0.658 = 0.342$$

Fisher's Exact Test: P-values

- To compute p-values, consider all 2×2 tables possible given the observed margins
- One-sided p-value: sum the probabilities of the observed table and all tables more extreme than the observed table in the direction of H_A
- Two-sided p-value: sum the probabilities of tables that are as likely as or less likely than the observed table, given the fixed margins

Fisher's Exact Test: P-values

- Most statistical software packages compute the p-value for Fisher's exact test. The tables in the text are difficult to use.
- SAS:

```
data;  
    input surgery \$ discharge \$ count;  
    datalines;  
        emergency dead 1  
        emergency alive 19  
        other dead 7  
        other alive 369  
    ;  
  
proc freq order=data;  
    tables surgery*discharge / nopct nocol;  
    exact fisher;  
    weight count;
```

Fisher's Exact Test: SAS Output

	surgery	discharge		
	Frequency			
Row	Pct	dead	alive	Total
emergenc		1	19	20
		5.00	95.00	
other		7	369	376
		1.86	98.14	
Total		8	388	396

Fisher's Exact Test

Cell (1,1) Frequency (F)	1
Left-sided Pr <= F	0.9434
Right-sided Pr >= F	0.3419
Table Probability (P)	0.2854
Two-sided Pr <= P	0.3419

Fisher's Exact Test: P-values

- R

```
> fisher.test(matrix(c(1,19,7,369),nrow=2),alternative="greater")
```

Fisher's Exact Test for Count Data

```
data: matrix(c(1, 19, 7, 369), nrow = 2)
p-value = 0.3419
alternative hypothesis: true odds ratio is greater than 1
```

```
> fisher.test(matrix(c(1,19,7,369),nrow=2))
```

Fisher's Exact Test for Count Data

```
data: matrix(c(1, 19, 7, 369), nrow = 2)
p-value = 0.3419
alternative hypothesis: true odds ratio is not equal to 1
```

Fisher's Exact Test: Example II

- Suppose another study yields

	Dead	Alive	
Emergency	2	23	25
Non-emergency	5	30	35
Total	7	53	60

- Null hypothesis

$$H_0 : \Pr[\text{dead} | \text{emergency}] = \Pr[\text{dead} | \text{non-emergency}]$$

$$H_0 : \pi_1 = \pi_2$$

Fisher's Exact Test: Example II cont.

- p-value computation

a	$\Pr[n_{11} = a]$	$H_A : \pi_1 > \pi_2$	$H_A : \pi_1 < \pi_2$	$H_A : \pi_1 \neq \pi_2$
0	0.017		+	+
1	0.105		+	+
2	0.252	+	+	+
3	0.312	+		
4	0.214	+		+
5	0.082	+		+
6	0.016	+		+
7	0.001	+		+

Fisher's Exact Test: Example II cont.

- Critical region for $H_A : \pi_1 > \pi_2$

$$C_{0.10} = \{n_{11} : n_{11} = 5, 6, \text{ or } 7\}$$

- Critical region for $H_A : \pi_1 < \pi_2$

$$C_{0.10} = \{n_{11} : n_{11} = 0\}$$

- Critical region for $H_A : \pi_1 \neq \pi_2$

$$C_{0.10} = \{n_{11} : n_{11} = 0, 6, \text{ or } 7\}$$

Fisher's Exact Test: Comments

- Justification/ramification of conditioning on margins
- Alternative: Barnard's test, more powerful for small sample sizes. Available in StatXact. R?

Comparing Two Proportions: Large Samples

- If n_1 and n_2 are large, we can use the normal distribution
- Let n_{i1} be the number of successes in the i^{th} sample;
 $i = 1, 2$
- Estimator of π_i is $p_i = n_{i1}/n_i$
- From the CLT, if n_i is large

$$p_i \sim N\left(\pi_i, \frac{\pi_i(1 - \pi_i)}{n_i}\right)$$

Comparing Two Proportions: Large Samples

- If samples are independent and π_i known for $i = 1, 2$, it follows

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0, 1)$$

- This approximation is good if $n_i\pi_i(1 - \pi_i) \geq 10$ for $i = 1, 2$

Comparing Two Proportions: Large Samples

- If samples are independent and π_i unknown for $i = 1, 2$, Slutsky/CLT imply

$$\frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$$

for sufficiently large n_1 and n_2

(rule of thumb: $n_i p_i (1 - p_i) \geq 10$ for $i = 1, 2$)

Comparing Two Proportions: Example

- A case-control study was conducted to investigate the association between oral contraceptive use and myocardial infarction
- Among 234 MI patients, 29 were OC users
- Among 1,742 non-MI patients, 135 were OC users
- Let π_1 denote the probability of OC use given a case (MI) and π_2 denote the probability of OC use given a control (no MI)

Comparing Two Proportions: Example cont.

- Hypotheses

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_A : \pi_1 \neq \pi_2$$

- Rejection region

$$C_{0.05} = \{|z| > 1.96\}$$

- Point estimates

$$p_1 = 29/234 = 0.124; \quad p_2 = 135/1742 = 0.078$$

- Test statistic

$$z = \frac{0.124 - 0.078 - 0}{\sqrt{\frac{(0.124)(0.876)}{234} + \frac{(0.078)(0.922)}{1742}}} = 2.42$$

Comparing Two Proportions: χ^2 Test

- Alternative test of $H_0 : \pi_1 = \pi_2$ is the χ^2 test
- Recall 2×2 table

	Success	Failure	
Sample 1	n_{11}	n_{12}	n_1
Sample 2	n_{21}	n_{22}	n_2
	m_1	m_2	N

- It can be shown that under H_0 , the statistic

$$X^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_1n_2m_1m_2} \sim \chi_1^2$$

- Critical region for $H_A : \pi_1 \neq \pi_2$

$$C_\alpha = \{X^2 : X^2 \geq \chi_{1,1-\alpha}^2\}$$

Comparing Two Proportions: χ^2 Test

- Also known as the “Pearson” chi-square statistic
- Equivalent form

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - E(n_{ij}))^2}{E(n_{ij})}$$

where $E(n_{ij}) = n_i m_j / N$

- We will see this again for $r \times c$ tables

Comparing Two Proportions: χ^2 Test

- OC-MI example:

	OC Users	Non-users	
MI Cases	29	205	234
Controls	135	1607	1742
	164	1812	1976

- Rejection region: $C_{0.05} = \{X^2 : X^2 > \chi^2_{1,0.95} = 3.84\}$
- Test statistic

$$X^2 = \frac{1976(29 \times 1607 - 135 \times 205)^2}{234 \times 1742 \times 1812 \times 164} = 5.84$$

χ^2 Test Example: SAS

```
proc freq order=data; tables patient*oc / norow nocol nopercent chisq;
```

Table of patient by oc

patient oc

	Frequency	yes	no	Total
mi		29	205	234
non-mi		135	1607	1742
Total		164	1812	1976

Statistics for Table of patient by oc

Statistic	DF	Value	Prob
<hr/>			
Chi-Square	1	5.8443	0.0156

χ^2 Test Example: R

```
> chisq.test(matrix(c(29,205,135,1607),nrow=2),correct=FALSE)
```

Pearson's Chi-squared test

```
data: matrix(c(29, 205, 135, 1607), nrow = 2)
X-squared = 5.8443, df = 1, p-value = 0.01563
```

```
> chisq.test(matrix(c(29,205,135,1607),nrow=2))
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: matrix(c(29, 205, 135, 1607), nrow = 2)
X-squared = 5.2501, df = 1, p-value = 0.02195
```

Comparing Two Proportions: χ^2 Test

- Note: $\sqrt{5.84} = 2.42$ and $\sqrt{3.84} = 1.96$
- Intuition: If $Z \sim N(0, 1)$, then $Z^2 \sim \chi_1^2$
- Indeed, for 2-sided tests, the χ^2 and Z test are approximately equivalent
- In fact, if we use

$$Z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where $p = (n_{11} + n_{21})/N$,

then exactly equivalent for two-sided H_A

Comparing Two Proportions: Summary

- For small samples, use Fisher's exact test
- For large samples and $H_A : \pi_1 \neq \pi_2$, use χ^2 or Z test, i.e.,

$$C_\alpha = \{X^2 : X^2 > \chi_{1,1-\alpha}^2\}$$

or $C_\alpha = \{z : |z| > z_{1-\alpha/2}\}$

- For large samples and $H_A : \pi_1 < \pi_2$ or $H_A : \pi_1 > \pi_2$, use Z test, i.e.,

$$C_\alpha = \{z : z < -z_{1-\alpha}\}$$

or $C_\alpha = \{z : z > z_{1-\alpha}\}$

Outline

- One sample binary outcome
- Two sample binary outcome
- Measures of association
 - Risk difference
 - Relative risk (risk ratio)
 - Odds ratio
- Confounding - Mantel-Haenszel
- Matching - McNemar

Measures of Association

- In epidemiologic studies, we often obtain 2×2 tables

	Disease	No disease	
Exposed	n_{11}	n_{12}	n_1
Unexposed	n_{21}	n_{22}	n_2
	m_1	m_2	N

- Source could be a cross-sectional, case-control, or prospective (cohort or clinical trial) study

Measures of Association: Estimands

- Let

$$\begin{aligned}\pi_1 &= \Pr[\text{ disease } | \text{ exposed }] \\ \text{and } \pi_2 &= \Pr[\text{ disease } | \text{ not exposed}]\end{aligned}$$

- Risk difference:

$$RD = \pi_1 - \pi_2$$

- Risk ratio (relative risk):

$$RR = \pi_1 / \pi_2$$

- Odds ratio (cross product ratio):

$$OR = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)}$$

Measures of Association: Estimands

- Independence or no association corresponds to

$$RR = 1 \text{ and } OR = 1$$

- $OR, RR \in [0, \infty)$
- $RR = 4$ implies an exposed person is 4 times as likely to have the disease as an unexposed person
- $OR = 4$ implies the odds of disease in the exposed is 4 times that in the unexposed

Measure of Association: Estimands

- Note

$$\text{OR/RR} = \left[\frac{\pi_1(1 - \pi_2)}{\pi_2(1 - \pi_1)} \right] / \left[\frac{\pi_1}{\pi_2} \right] = \frac{1 - \pi_2}{1 - \pi_1}$$

- If disease rare,

$$1 - \pi_1 \approx 1 - \pi_2 \approx 1$$

- In this case, $\text{OR} \approx \text{RR}$; this is important in case-control studies

- Rule of thumb:

$\pi_1, \pi_2 \leq 0.05$ (text page 165);

$\pi_1, \pi_2 \leq 0.10$ (Rosner, 1995, page 368);

requires external knowledge

Measures of Association: Estimators

- Risk difference:

$$\widehat{RD} = p_1 - p_2 = (n_{11}/n_1) - (n_{21}/n_2)$$

- Relative risk:

$$\widehat{RR} = p_1/p_2 = (n_{11}/n_1)/(n_{21}/n_2)$$

- Odds ratio:

$$\widehat{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

Estimating RR in Case-Control Studies

- In case-control studies, \widehat{RR} should not be used to estimate RR. Why?
- Intuitively, RR describes $\Pr[D^+|E^+]$ and $\Pr[D^+|E^-]$, while case-control studies provide information about $\Pr[E^+|D^+]$ and $\Pr[E^+|D^-]$

Estimating RR in Case-Control Studies

- Formally: Suppose the joint distribution of exposure and disease in the population is denoted by

		Disease	No disease	
		π_{11}	π_{12}	$\pi_{1\cdot}$
		π_{21}	π_{22}	$\pi_{2\cdot}$
		$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	

Estimating RR in Case-Control Studies

- Sample m_1 individuals with disease and m_2 without disease.
- The expected numbers of observations are

		Disease	No disease
Exposed		$\frac{m_1\pi_{11}}{\pi_{.1}}$	$\frac{m_2\pi_{12}}{\pi_{.2}}$
		$\frac{m_1\pi_{21}}{\pi_{.1}}$	$\frac{m_2\pi_{22}}{\pi_{.2}}$
		m_1	m_2

Estimating RR in Case-Control Studies

- Therefore

$$\begin{aligned}\widehat{\text{RR}} &\approx \frac{\left(\frac{m_1\pi_{11}}{\pi_{.1}}\right) / \left(\frac{m_1\pi_{11}}{\pi_{.1}} + \frac{m_2\pi_{12}}{\pi_{.2}}\right)}{\left(\frac{m_1\pi_{21}}{\pi_{.1}}\right) / \left(\frac{m_1\pi_{21}}{\pi_{.1}} + \frac{m_2\pi_{22}}{\pi_{.2}}\right)} \\ &= \frac{\pi_{11} \times \left(\frac{m_1\pi_{21}}{\pi_{.1}} + \frac{m_2\pi_{22}}{\pi_{.2}}\right)}{\pi_{21} \times \left(\frac{m_1\pi_{11}}{\pi_{.1}} + \frac{m_2\pi_{12}}{\pi_{.2}}\right)} \\ &= \frac{\frac{m_1\pi_{11}\pi_{21}}{\pi_{.1}} + \frac{m_2\pi_{11}\pi_{22}}{\pi_{.2}}}{\frac{m_1\pi_{11}\pi_{21}}{\pi_{.1}} + \frac{m_2\pi_{12}\pi_{21}}{\pi_{.2}}}\end{aligned}$$

- This depends on the choice of m_1 and m_2 ;
for instance, $\widehat{\text{RR}} \rightarrow 1$ as $m_1 \rightarrow \infty$ for fixed m_2

Estimating RR in Case-Control Studies

- On the other hand, we would expect

$$\widehat{OR} \approx \frac{\left(\frac{m_1\pi_{11}}{\pi_{.1}}\right) / \left(\frac{m_2\pi_{12}}{\pi_{.2}}\right)}{\left(\frac{m_1\pi_{21}}{\pi_{.1}}\right) / \left(\frac{m_2\pi_{22}}{\pi_{.2}}\right)}$$
$$= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}$$

- For a rare disease, π_{11} and π_{21} are both small, so

$$\frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \approx \frac{\pi_{11}/(\pi_{11} + \pi_{12})}{\pi_{21}/(\pi_{21} + \pi_{22})}$$

Thus $\widehat{OR} \approx RR$ in this case.

Estimating OR in Case-Control Studies

- Intuitively, why does \widehat{OR} estimate OR in a case-control study?

$$\begin{aligned} OR &= \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \\ &= \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\frac{\pi_{11}}{\pi_{11} + \pi_{21}} / \frac{\pi_{21}}{\pi_{11} + \pi_{21}}}{\frac{\pi_{12}}{\pi_{12} + \pi_{22}} / \frac{\pi_{22}}{\pi_{12} + \pi_{22}}} \\ &= \frac{\omega_1/(1 - \omega_1)}{\omega_2/(1 - \omega_2)} \end{aligned}$$

where

$$\omega_1 = \pi_{11}/(\pi_{11} + \pi_{21}) = \Pr[E+ | D+]$$

$$\omega_2 = \pi_{12}/(\pi_{12} + \pi_{22}) = \Pr[E+ | D-]$$

Measures of Association: RD

- Similarly, \widehat{RD} should not be used to estimate RD in case-control studies
- For prospective or cross-sectional studies, a $100(1 - \alpha)\%$ CI for RD is given by

$$p_1 - p_2 \pm z_{1-\alpha/2} \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

when n_1 and n_2 are sufficiently large

Measures of Association: RR

- It can be shown that

$$\widehat{\text{Var}}(\log(\widehat{RR})) = \frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}$$

and

$$\log(\widehat{RR}) \sim N(\log(RR), \text{Var}(\log(RR)))$$

- Therefore a $100(1 - \alpha)\%$ CI for $\log(RR)$ is

$$\log(p_1/p_2) \pm z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}}$$

Measures of Association: RR

- Thus

$$\text{CI}_{\text{lower}} = \frac{p_1}{p_2} \exp \left\{ -z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}} \right\}$$

$$\text{CI}_{\text{upper}} = \frac{p_1}{p_2} \exp \left\{ z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}} \right\}$$

- In a prospective or cross-sectional study, these CIs are recommended when $n_i p_i (1 - p_i) \geq 5$ for $i = 1, 2$ where p_1 and p_2 are the sample proportions with the disease given exposed and unexposed, respectively
- See Rosner (1995) page 364

Measures of Association: Example

- In a study of the relationship between obesity and asthma, a cohort of 3,792 children free of asthma were followed for 5 years

	Asthma	No asthma	
Obese	36	154	190
Not obese	252	3350	3602
	288	3504	3792

Measures of Association: Example cont.

- Null hypothesis

$$H_0 : \Pr[\text{asthma} \mid \text{obese}] = \Pr[\text{asthma} \mid \text{not obese}]$$

$$H_0 : \pi_1 = \pi_2$$

Equivalently:

$$H_0 : \pi_1 - \pi_2 = 0 \quad \text{or} \quad H_0 : \pi_1 / \pi_2 = 1$$

- Rejection region

$$C_{0.05} = \{X^2 > 3.84\}$$

- Test statistic

$$X^2 = \frac{(3792)(36 \times 3350 - 252 \times 154)^2}{3602 \times 190 \times 288 \times 3504} = 36.73$$

Measures of Association: Example cont.

- Point estimate of RD

$$\begin{aligned}\widehat{\text{RD}} &= p_1 - p_2 \\ &= 36/190 - 252/3602 \\ &= 0.189 - 0.070 \\ &= 0.12\end{aligned}$$

Interpretation: we estimate that obese children have a 12 percentage point greater chance of developing asthma within 5 years than non-obese children

- 95% CI: (0.063, 0.176)

Measures of Association: Example cont.

- Point estimate of RR

$$\widehat{RR} = 0.189/0.070 = 2.7$$

Interpretation: we estimate that obese children are 2.7 times more likely to develop asthma within 5 years than non-obese children

- 95% CI for RR:

$$2.7 \exp \left\{ \pm 1.96 \sqrt{\frac{154}{36(190)} + \frac{3350}{252(3602)}} \right\} = (1.97, 3.72)$$

Measures of Association: SAS Code/Output

```
data;  
    input asthma \$ obese \$ count;  
    datalines;  
    yes yes 36  
    yes no 252  
    no yes 154  
    no no 3350  
    ;  
proc freq order=data;  
    tables obese*asthma / norow nocol nopercent relrisk riskdiff;  
    weight count;
```

Table of obese by asthma

obese	asthma			Total		
	Frequency	yes	no			
yes		36		154		190
no		252		3350		3602
Total		288		3504		3792

Measures of Association: SAS Code/Output

Statistics for Table of obese by asthma

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits	

Row 1	0.1895	0.0284	0.1338	0.2452
Row 2	0.0700	0.0043	0.0616	0.0783
Total	0.0759	0.0043	0.0675	0.0844
Difference	0.1195	0.0287	0.0632	0.1759

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits

Case-Control (Odds Ratio)	3.1076	2.1151 4.5659
Cohort (Col1 Risk)	2.7083	1.9720 3.7195
Cohort (Col2 Risk)	0.8715	0.8131 0.9341

Measures of Association: OR

- Can show

$$\widehat{\text{Var}}(\log(\widehat{\text{OR}})) = \frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}$$

and

$$\log(\widehat{\text{OR}}) \sim N(\log(\text{OR}), \text{Var}(\log(\text{OR})))$$

(Woolf, 1955)

- Thus for large n , a $100(1 - \alpha)\%$ CI is

$$\widehat{\text{OR}} \exp \left\{ \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}} \right\}$$

Measures of Association: OR

- In a prospective or cross-sectional study, Woolf CIs are recommended when

$$n_i p_i (1 - p_i) \geq 5$$

for $i = 1, 2$ where p_1 and p_2 are the sample proportions with disease given exposed and unexposed, respectively

- In a case-control study, Woolf CIs are recommended when

$$m_i p_i^* (1 - p_i^*) \geq 5$$

for $i = 1, 2$ where p_1^* and p_2^* are the sample proportions exposed among cases and controls, respectively

- See Rosner (1995) page 369

Measures of Association: OR

- Recall the oral contraceptive use and MI example:

	OC Users	Non-users	
MI Cases	29	205	234
Controls	135	1607	1742
	164	1812	1976

- Point estimate

$$\widehat{OR} = \frac{29 \times 1607}{205 \times 135} = 1.68$$

- 95% CI

$$1.684 \exp \left\{ \pm 1.96 \sqrt{\frac{1}{29} + \frac{1}{205} + \frac{1}{135} + \frac{1}{1607}} \right\} = (1.10, 2.58)$$

Measures of Association: OR

- SAS output:

Table of patient by oc

patient	oc			
	Frequency	yes	no	Total
mi		29	205	234
non-mi		135	1607	1742
Total		164	1812	1976

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits
<hr/>		
Case-Control (Odds Ratio)	1.6839	1.0991 2.5800
Cohort (Col1 Risk)	1.5992	1.0967 2.3320
Cohort (Col2 Risk)	0.9497	0.9033 0.9984

Measures of Association: OR

- R

```
> # First need to install the "epitools" package  
> library(epitools)  
  
> # Rows should be the exposures, columns the case status  
> # Unexposed controls should be in top left cell  
> example <-  
    array(c(1607,135,205,29),  
          dim = c(2, 2),  
          dimnames = list(OC = c("Non-user", "User"),  
                         MI = c("Control", "Case")))
```

Measures of Association: OR

- R

```
> oddsratio.wald(example)
```

```
$data
```

```
MI
```

OC	Control	Case	Total
Non-user	1607	205	1812
User	135	29	164
Total	1742	234	1976

```
$measure
```

```
odds ratio with 95% C.I.
```

OC	estimate	lower	upper
Non-user	1.000000	NA	NA
User	1.683939	1.099069	2.580045

```
$p.value
```

```
two-sided
```

OC	midp.exact	fisher.exact	chi.square
Non-user	NA	NA	NA
User	0.02158681	0.02228029	0.01562785

Confounding

- *Confounding:* A confounding variable is a variable that is associated with both the disease and the exposure.
- Such a variable may bias the measured association between exposure and disease
- A confounding variable may mask a true disease-exposure association or may cause the observed association to be too large

Confounding: Example

- Malaria and gender (case-control study)

	Malaria	No malaria	
Males	88	68	156
Females	62	82	144
	150	150	300

- Null hypothesis

$$H_0 : \pi_1 = \pi_2 \Leftrightarrow H_0 : \text{OR} = 1$$

- $\widehat{\text{OR}} = 1.71$; $X^2 = 5.34$ ($p = 0.02$)
- However, men work outdoors more than women

Confounding: Example cont.

- Stratified analysis
- Outdoor occupation $\widehat{OR} = 1.06$

	Malaria	No malaria	
Males	53	15	68
Females	10	3	13
	63	18	81

- Indoor occupation $\widehat{OR} = 1.00$

	Malaria	No malaria	
Males	35	53	88
Females	52	79	131
	87	132	219

Confounding: Mantel-Haenszel

- Adjust for possible confounding by stratification and combining 2×2 tables.
- For each stratum, $j = 1, 2, \dots, S$, we have

	Disease	No disease	
Exposed	n_{11j}	n_{12j}	n_{1j}
Unexposed	n_{21j}	n_{22j}	n_{2j}
	m_{1j}	m_{2j}	N_j

- Recall that if the margins $(m_{1j}, m_{2j}, n_{1j}, n_{2j})$ are fixed, n_{11j} follows the hypergeometric distribution

Confounding: Mantel-Haenszel

- Thus

$$E(n_{11j}) = \frac{n_{1j}m_{1j}}{N_j}$$

and

$$\text{Var}(n_{11j}) = \frac{n_{1j}n_{2j}m_{1j}m_{2j}}{N_j^2(N_j - 1)}$$

- Let

$$O_j = n_{11j}; \quad E_j = E(n_{11j}); \quad V_j = \text{Var}(n_{11j})$$

and

$$O = \sum_{j=1}^S O_j; \quad E = \sum_{j=1}^S E_j; \quad V = \sum_{j=1}^S V_j;$$

Confounding: Mantel-Haenszel

- The Mantel-Haenszel statistic is given by

$$X_{\text{MH}}^2 = \frac{(|O - E| - 0.5)^2}{V}$$

- Under $H_0 : \text{OR} = 1$ within strata, $X_{\text{MH}}^2 \sim \chi_1^2$

$$C_\alpha = \{X_{\text{MH}}^2 : X_{\text{MH}}^2 > \chi_{1,1-\alpha}^2\}$$

- X_{MH} has power against the alternative hypothesis of consistent patterns of association; it has low power for detecting association in opposite directions. However, it always preserves type I error (Stokes, Davis, Koch 1995)

Confounding: Mantel-Haenszel

- Assuming homogeneous OR across strata, we can also use the MH approach to estimate the overall or common OR
- MH estimator of OR

$$\widehat{\text{OR}}_{\text{MH}} = \frac{\sum_{j=1}^S n_{11j} n_{22j} / N_j}{\sum_{j=1}^S n_{12j} n_{21j} / N_j}$$

Confounding: Mantel-Haenszel

- Let

$$P_j = (n_{11j} + n_{22j})/N_j; \quad Q_j = (n_{12j} + n_{21j})/N_j$$

$$R_j = (n_{11j} n_{22j})/N_j; \quad W_j = (n_{12j} n_{21j})/N_j$$

- Then $\text{Var}(\log(\widehat{\text{OR}}_{\text{MH}}))$ is

$$\frac{\sum_j P_j R_j}{2(\sum_j R_j)^2} + \frac{\sum_j (P_j W_j + Q_j R_j)}{2(\sum_j R_j)(\sum_j W_j)} + \frac{\sum_j Q_j W_j}{2(\sum_j W_j)^2}$$

- A $100(1 - \alpha)\%$ CI is

$$\widehat{\text{OR}}_{\text{MH}} \exp \left\{ \pm z_{1-\alpha/2} \sqrt{\text{Var}(\log(\widehat{\text{OR}}_{\text{MH}}))} \right\}$$

- Robins, Breslow, Greenland (Biometrics, 1986);

See Rosner 1995 p 410

Confounding: Malaria Example Revisited

- Unstratified: $X^2 = 5.34$
- Outdoor $\widehat{OR} = 1.06$; indoor $\widehat{OR} = 1.00$
- Outdoor:

$$O_1 = 53; \quad E_1 = \frac{68 \times 63}{81} = 52.889;$$

$$V_1 = \frac{68 \times 13 \times 63 \times 18}{81^2 \times 80} = 1.9099$$

- Indoor:

$$O_2 = 35; \quad E_2 = 34.9589; \quad V_2 = 12.6620$$

Confounding: Malaria Example cont.

- MH test statistic

$$X_{\text{MH}}^2 = \frac{(|(53 + 35) - (52.889 + 34.9589)| - 0.5)^2}{1.9099 + 12.6620}$$
$$= 0.008$$

without continuity correction $X_{\text{MH}}^2 = 0.0016$

Confounding: Malaria Example Using SAS

```
** Note that the confounder is the first variable;  
** listed in the tables statement;
```

```
proc freq order=data;  
  tables job*gender*malaria / cmh;  
  weight count;
```

Summary Statistics for gender by malaria
Controlling for job

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob

1	Nonzero Correlation	1	0.0016	0.9682
2	Row Mean Scores Differ	1	0.0016	0.9682
3	General Association	1	0.0016	0.9682

Confounding: Malaria Example using R

```
example <- array(c(53,10,15,3,35,52,53,79),  
                  dim = c(2, 2),  
                  dimnames = list(Gender = c("Male", "Female"),  
                                  Malaria = c("Yes", "No"),  
                                  Job = c("Outdoors", "Indoors")))  
  
> mantelhaen.test(example)
```

Mantel-Haenszel chi-squared test without continuity correction

```
data: example  
Mantel-Haenszel X-squared = 0.0016, df = 1, p-value = 0.9682  
alternative hypothesis: true common odds ratio is not equal to 1  
95 percent confidence interval:  
 0.6041733 1.6902399  
sample estimates:  
common odds ratio  
 1.010543
```

Matched or Paired Observations

- In some studies, subjects occur naturally in pairs or matches; e.g., twins or a matched case-control design
- If we want to compare binary responses in matched pairs, the assumption of independence is violated
- The data are of the form (Y_{i1}, Y_{i2}) , where $Y_{ij} = 1$ if exposed and $= 0$ if unexposed; $i = 1, 2, \dots, n$; $j = 1, 2$

		D^+		n
		$Y_{i1} = 1$	$Y_{i1} = 0$	
D^-	$Y_{i2} = 1$	n_{11}	n_{12}	
	$Y_{i2} = 0$	n_{21}	n_{22}	
				n

Matched or Paired Observations

- Note

$$\Pr[Y_{i1} = 1] = \Pr[Y_{i1} = 1, Y_{i2} = 1] + \Pr[Y_{i1} = 1, Y_{i2} = 0]$$

and

$$\Pr[Y_{i2} = 1] = \Pr[Y_{i1} = 1, Y_{i2} = 1] + \Pr[Y_{i1} = 0, Y_{i2} = 1]$$

- Therefore

$$\pi_1 - \pi_2 = \Pr[Y_{i1} = 1] - \Pr[Y_{i2} = 1]$$

$$= \Pr[Y_{i1} = 1, Y_{i2} = 0] - \Pr[Y_{i1} = 0, Y_{i2} = 1]$$

Matched or Paired Observations

- Hypotheses

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_A : \pi_1 \neq \pi_2$$

- McNemar's test statistic

$$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

- Under H_0 , $M \sim \chi_1^2$ if $n_{12} + n_{21}$ is sufficiently large
(i.e. ≥ 30)

$$\begin{aligned} C_\alpha &= \{M : M > \chi_{1,1-\alpha}^2\} \\ p &= \Pr[\chi_1^2 \geq m] \end{aligned}$$

Matched/Paired Observations: Example

- A case-control study was conducted to investigate the association between cytomegalovirus (CMV) and atherosclerosis
- Study participants with atherosclerosis, as measured by ultrasound of the carotid artery, were matched with persons without atherosclerosis, matching on age, sex, ethnicity, geographic site, and date of ultrasound
- Cytomegalovirus antibodies were measured in each person

Matched/Paired Observations: Example cont.

		Cases	
		CMV+	CMV-
Controls	CMV+	214	42
	CMV-	65	19

- McNemar's test statistic

$$M = \frac{(42 - 65)^2}{42 + 65} = 4.94$$

- Reject $H_0 : \pi_1 = \pi_2$ for $\alpha = 0.05$;

$$p = \Pr[\chi_1^2 \geq 4.94] = 0.026$$

Matched or Paired Observations

- The χ^2 approximation for McNemar's test is adequate if $n_{12} + n_{21} \geq 30$
- For smaller samples, can compute the exact p-value
- Key: recognize this as a one sample binomial test
- Let $c = n_{12} + n_{21}$. If $n_{12} < c/2$, then

$$p = 2 \sum_{k=0}^{n_{12}} \binom{c}{k} 2^{-c}$$

otherwise

$$p = 2 \sum_{k=n_{12}}^c \binom{c}{k} 2^{-c} = 2 \sum_{k=0}^{n_{21}} \binom{c}{k} 2^{-c}$$

Matched/Paired Observations: Example II

- Suppose we want to compare 2 lotions for the treatment of poison ivy
- Persons with poison ivy on both arms are selected for the study
- One arm is randomly assigned to receive lotion 1, while the other is treated with lotion 2

		Lotion 1	
		Relief	No relief
Lotion 2	Relief	11	6
	No relief	10	24

Matched/Paired Observations: Example II cont.

- Let $\pi_i = \Pr(\text{itching relief using lotion } i)$

$$H_0 : \pi_1 = \pi_2 \text{ vs. } H_A : \pi_1 \neq \pi_2$$

- Exact p-value

$$p = 2 \sum_{k=0}^6 \binom{16}{k} 2^{-16} = 2 \times 0.2272 = 0.4544$$

- Do not reject H_0

$$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(6 - 10)^2}{6 + 10} = 1$$

- R: `mcnemar.test()`

Matched/Paired Observations: Example II cont.

```
proc freq order=data;
  tables lotion2*lotion1 / norow nocol nopercnt;
  exact agree; weight count;
```

The FREQ Procedure

Table of lotion2 by lotion1

		lotion1		Total
		relief	norelief	
lotion2	relief	11	6	17
	norelief	10	24	34
Total		21	30	51

Statistics for Table of lotion2 by lotion1

McNemar's Test

Statistic (S) 1.0000
DF 1
Asymptotic Pr > S 0.3173
Exact Pr >= S 0.4545

McNemar's Test

- *Marginal homogeneity*

$$H_0 : \Pr[Y_{i1} = 1] = \Pr[Y_{i2} = 1]$$

- This is a test of association with a risk factor, not a test for agreement between the members of a pair; consider

		Rater 1	
		+	-
Rater 2	+	0	65
	-	65	0

for these data: $M = 0$; $p = 1$

- We'll look at a measure of agreement (kappa statistic) later in the semester

Matched or Paired Observations

- Odds ratio for matched data

$$\widehat{OR}_M = n_{21}/n_{12}$$

this is just \widehat{OR}_{MH} with a stratum for each matched pair

- Confidence interval obtained by starting on the log scale

$$\widehat{\text{Var}}(\ln(\widehat{OR}_M)) \approx \frac{1}{n_{12}} + \frac{1}{n_{21}}$$

- For $n_{12} + n_{21} \geq 30$, an approximate $100(1 - \alpha)\%$ CI

$$\exp \left(\ln(\widehat{OR}_M) \pm z_{1-\alpha/2} \sqrt{\widehat{\text{Var}}(\ln(\widehat{OR}_M))} \right)$$

Matched or Paired Observations: CMV Example

- Odds ratio estimate

$$\widehat{OR}_M = 65/42 = 1.55$$

- Corresponding estimate of variance of $\ln(\widehat{OR}_M)$

$$\widehat{\text{Var}}(\ln(\widehat{OR}_M)) = \frac{1}{65} + \frac{1}{42} = 0.0392$$

- Approximate 95% CI on the log scale

$$\ln(1.55) \pm 1.96 \times \sqrt{0.0392} = (0.0502, 0.8263)$$

- So an approximate 95% CI on the original scale is

$$(e^{0.0502}, e^{0.8263}) = (1.05, 2.28)$$

BIOS 662 Fall 2016

Categorical Data: Contingency Tables

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Contingency Tables

- Two-way $(r \times c)$ contingency table:

		j			
		1	2	\dots	c
1		n_{11}	n_{12}	\dots	n_{1c}
2		n_{21}	n_{22}	\dots	n_{2c}
i	:	:	:	:	:
r		n_{r1}	n_{r2}	\dots	n_{rc}

- Notation:

$$n_{i\cdot} = \sum_{j=1}^c n_{ij} \quad n_{\cdot j} = \sum_{i=1}^r n_{ij}$$

Contingency Tables

- Two scenarios where $r \times c$ tables arise
 1. Sample from a population and measure two characteristics, say X and Y

$$\Pr[X = i, Y = j] = \pi_{ij}; \quad \sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = 1$$

2. Each row corresponds to a sample from a different population

$$\sum_{j=1}^c \pi_{ij} = 1$$

Contingency Table: Example I

- A survey of physicians asked about the size of the community in which they were reared and the size of the community in which they practice

Reared	Practice				Total
	<5K	5-49K	50-99K	$\geq 100K$	
<5K	40	38	32	37	147
5-49K	26	42	35	33	136
50-99K	24	26	34	31	115
$\geq 100K$	30	39	53	60	182
	120	145	154	161	580

Contingency Table: Example II

- A case-control study of women was conducted to investigate the relationship between age at which they first gave birth and breast cancer

		Age at first childbirth					Total
		<20	20-24	25-29	30-34	≥ 35	
Case	320	1206	1011	463	220	3220	3220
	1422	4432	2893	1092	406	10245	
		1742	5638	3904	1555	626	13465

Contingency Tables

- For the physicians example, H_0 : size of community in which practice is independent of size of community in which reared

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$$

- Test of independence, $X \perp Y$

$$\Pr[X = i, Y = j] = \Pr[X = i] \Pr[Y = j]$$

for $i = 1, \dots, r; j = 1, \dots, c$

Contingency Tables

- For the breast cancer example, H_0 : distribution of age at first childbirth is the same for cases and controls

$$H_0 : \pi_{ij} = \pi_{i'j}; \quad j = 1, 2, \dots, c$$

- Test of homogeneity/association

Test of Independence or No Association

- Under either H_0 , the estimated expected frequency in the (i, j) cell is

$$E_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{N}$$

- Consider the breast cancer example
 - The overall proportion of women <20 is
$$\frac{n_{11} + n_{21}}{N} = \frac{n_{\cdot 1}}{N}$$
 - There are $n_{1\cdot}$ cases, so if H_0 is true we would expect

$$E_{11} = n_{1\cdot} \cdot \frac{n_{\cdot 1}}{N} = \frac{n_{1\cdot}n_{\cdot 1}}{N}$$

cases to be <20 years old

Test of Independence or Association

- Under H_0 , the expected frequency in the (i, j) cell is

$$E_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{N}$$

- Let

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

that is,

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/N)^2}{n_{i\cdot}n_{\cdot j}/N}$$

Test of Independence

- Under H_0 ,

$$X^2 \sim \chi^2_{(r-1)(c-1)}$$

- Physicians example:

$$(r - 1)(c - 1) = 3 \times 3 = 9$$

$$C_{0.05} = \{X^2 : X^2 > \chi^2_{9,0.95} = 16.92\}$$

Physicians Example

- Expected values

Reared	Practice				Total
	<5K	5-49K	50-99K	$\geq 100K$	
<5K	30.4	36.8	39.0	40.8	147
5-49K	28.1	34.0	36.1	37.8	136
50-99K	23.8	28.8	30.5	31.9	115
$\geq 100K$	37.7	45.5	48.3	50.5	182
	120	145	154	161	580

Physicians Example cont.

- Calculate the test statistic

$$X^2 = \frac{(40 - 30.4)^2}{30.4} + \frac{(38 - 36.8)^2}{36.8} + \cdots + \frac{(60 - 50.5)^2}{50.5}$$
$$= 12.81$$

- Do not reject H_0 .
- There is insufficient evidence to conclude that the size of the community in which practice and that of the community in which reared are dependent; the data are consistent with the null hypothesis that size of community in which practice and that in which reared are independent

Breast Cancer Example

- Underlying probabilities

		Age at first childbirth					Total
		<20	20-24	25-29	30-34	≥ 35	
Case	Case	π_{11}	π_{12}	π_{13}	π_{14}	π_{15}	1
	Control	π_{21}	π_{22}	π_{23}	π_{24}	π_{25}	1

- Null hypothesis

$$H_0 : \pi_{1j} = \pi_{2j} \text{ for } j = 1, 2, 3, 4, 5$$

- We can use the same statistic

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi_{(c-1)}^2$$

Breast Cancer Example cont.

- Expected frequencies

		Age at first childbirth					Total
		<20	20-24	25-29	30-34	≥ 35	
Case	416.6	1348.3	933.6	371.9	149.7	3220	3220
	1325.4	4289.7	2970.4	1183.1	476.3	10245	
		1742	5638	3904	1555	626	13465

Breast Cancer Example cont.

- Test statistic

$$X^2 = \frac{(320 - 416.6)^2}{416.6} + \cdots + \frac{(406 - 476.3)^2}{476.3} = 130.3$$

- Rejection region

$$C_{0.05} = \{X^2 : X^2 > \chi^2_{4,0.95} = 9.49\}$$

- Reject H_0
- Conclude that the age distributions are not the same

Asymptotic Approximation

- Note that the χ^2 distribution for X^2 is an approximation
- The approximation works well if $E_{ij} \geq 5$ for all i, j
- If $E_{ij} < 5$, a generalization of Fisher's exact test can be employed or categories combined

Test of Independence

- For $r = c = 2$, one can show that

$$\begin{aligned} X^2 &= \sum_{i=1}^2 \sum_{j=1}^2 \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/N)^2}{n_{i\cdot}n_{\cdot j}/N} \\ &= \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{\cdot 1}n_{2\cdot}n_{\cdot 2}} \end{aligned}$$

- This is the *Pearson chi-square statistic* we saw for comparing two proportions

Test for Trend

- Consider a $2 \times c$ table
- The χ^2 test for homogeneity does not tell us how the probabilities differ
- Rather, it tests just whether they differ
- If the categories of the column variable are ordered, a more powerful test is possible

Test for Trend

- Suppose the columns correspond to ordered levels of an exposure
- Suppose the rows correspond to disease (yes/no)
- We are interested in detecting alternatives where the probability of disease increases (or decreases) with exposure level
- That is, we are looking for a monotonic dose-response type of relationship

Test for Trend: Example

- Example 7.3 in the text: Risk of catheter-related infection and the duration of catheterization

		Duration (days)			
		1	2	3	4+
Culture	Positive	1	5	5	14
		46	64	39	76
Total		47	69	44	90

- Let ρ_j denote the conditional probability of being in row 1 given in column j
- For this catheter example, ρ_j is the probability of a positive culture given in the j^{th} duration category

Test for Trend

- Test

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_c$$

versus

$$H_A : \rho_1 \leq \rho_2 \leq \cdots \leq \rho_c$$

with at least one strict inequality, or

$$H_A : \rho_1 \geq \rho_2 \geq \cdots \geq \rho_c$$

with at least one strict inequality

Test for Trend

- Numerical scores must be assigned to the categories:

$$x_j : j = 1, 2, \dots, c$$

- Example: $x_j = j$
- In the breast cancer example, use the mid-range of age categories

Test for Trend

- Let

$$[n_1x] \equiv \sum_{j=1}^c n_{1j}x_j - \frac{n_{1\cdot} \sum_{j=1}^c n_{\cdot j}x_j}{N}$$

$$[x^2] \equiv \sum_{j=1}^c n_{\cdot j}x_j^2 - \frac{(\sum_{j=1}^c n_{\cdot j}x_j)^2}{N}$$

and

$$p \equiv \frac{n_{1\cdot}}{N}$$

- Then the chi-square test for trend (text, p 215) is

$$X_{\text{trend}}^2 \equiv \frac{[n_1x]^2}{[x^2]p(1-p)}$$

Test for Trend

- Huh?
- Compute the average score in row 1:

$$\bar{x} \equiv \sum_{j=1}^c \frac{n_{1j}x_j}{n_{1\cdot}}$$

- Compute the finite-sample expected value under the null:

$$E(\bar{x}) \equiv E(x) \equiv \sum_{j=1}^c \frac{n_{\cdot j}x_j}{N}$$

Test for Trend

- Compute the finite-sample variance:

$$\text{Var}(\bar{x}) \equiv \left(\frac{1-f}{n_1} \right) \left[E(x^2) - E(x)^2 \right]$$

where $f = n_1./N$ is the sampling fraction and

$$E(x^2) \equiv \sum_{j=1}^c \frac{n_{\cdot j} x_j^2}{N}$$

- Then the chi-square test for trend can equivalently be written as

$$X_{\text{trend}}^2 = \frac{(\bar{x} - E(\bar{x}))^2}{\text{Var}(\bar{x})}$$

Test for Trend

- Under H_0 ,

$$X_{\text{trend}}^2 \sim \chi_1^2$$

$$C_\alpha = \{X^2 : X^2 > \chi_{1,1-\alpha}^2\}$$

$$p = \Pr[\chi_1^2 > x^2]$$

- Note that here χ^2 has 1 degree of freedom regardless of c

Test for Trend: Example

- Catheter example

Culture	Duration			
	1	2	3	4+
Positive	1	5	5	14
Negative	46	64	39	76
Total	47	69	44	90

- $X^2_{\text{trend}} = 6.98$; $p = 0.008$; reject H_0 and conclude that the probability of a positive culture increases with duration of catheterization

Test for Trend: R

```
# Also if prop.trend.test(c(1,5,5,14),c(47,69,44,90),c(1,2,3,4))
# or      prop.trend.test(c(1,5,5,14),c(47,69,44,90),c(4,3,2,1))
> prop.trend.test(c(1,5,5,14),c(47,69,44,90))
```

Chi-squared Test for Trend in Proportions

```
data: c(1, 5, 5, 14) out of c(47, 69, 44, 90) ,
using scores: 1 2 3 4
X-squared = 6.9764, df = 1, p-value = 0.008259
```

```
> prop.trend.test(c(1,5,5,14),c(47,69,44,90),c(1,2,3,6))
```

Chi-squared Test for Trend in Proportions

```
data: c(1, 5, 5, 14) out of c(47, 69, 44, 90) ,
using scores: 1 2 3 6
X-squared = 6.4248, df = 1, p-value = 0.01125
```

Test for Trend: SAS

```
data trend;  input culture $1-8 duration count;
cards;
positive 1 1
positive 2 5
positive 3 5
positive 4 14
negative 1 46
negative 2 64
negative 3 39
negative 4 76
;

proc freq data=trend order=data;
  tables culture*duration / chisq nopct norow trend;
  weight count;
```

		culture duration				Total		
		Frequency	Col Pct	1	2	3	4	
		-----+-----+-----+-----+	-----+-----+-----+-----+	-----+-----+-----+-----+	-----+-----+-----+-----+	-----+-----+-----+-----+	-----+-----+-----+-----+	-----+-----+-----+-----+
positive		1	5	5	14	25		
		2.13	7.25	11.36	15.56			
negative		46	64	39	76	225		
		97.87	92.75	88.64	84.44			
Total		47	69	44	90	250		

Test for Trend: SAS, cont.

Statistics for Table of culture by duration

Statistic	DF	Value	Prob

Chi-Square	3	6.9951	0.0721
Mantel-Haenszel Chi-Square	1	6.9485	0.0084

WARNING: 25% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

Cochran-Armitage Trend Test

Statistic (Z)	2.6413
One-sided Pr > Z	0.0041
Two-sided Pr > Z	0.0083

χ^2 Test of Goodness of Fit

- Goal: Assess how well a particular model fits the data
- General form

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi_{\text{df}}^2$$

- E_i computed under H_0
- Rejecting the null implies the model does not provide an adequate fit to the data

χ^2 Test of Goodness of Fit

- Multinomial: Generalization of binomial from 2 to K categories
- Suppose there are n independent trials, each with K possible outcomes having probabilities π_1, \dots, π_K
- Let n_i be the number of trials having outcome i , $i = 1, \dots, K$, such that

$$n = \sum_{i=1}^K n_i$$

- Then

$$E(n_i) = n\pi_i \text{ and } \text{Var}(n_i) = n\pi_i(1 - \pi_i)$$

χ^2 GOF: Genetics Example

- Example: Mendelian genetics hypothesizes a particular genotype should occur in the proportions
 $1 : 2 : 1$ (dominant, heterozygous, recessive)
- $H_0 : \pi_1 = 0.25, \pi_2 = 0.5, \pi_3 = 0.25$
- H_A : at least one of the equalities is false (in which case at least two must be false)
- Suppose in a study the genotypes have frequencies
 $n_1 = 21, n_2 = 62, n_3 = 17$

χ^2 GOF: Genetics Example cont.

- When the expected values are known, in general

$$X^2 = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{K-1}$$

$K = 3$ for the genetics example

- Under H_0 ,

$$E_1 = 0.25 \times 100 = 25$$

$$E_2 = 50$$

$$E_3 = 25$$

χ^2 GOF: Genetics Example cont.

- Thus

$$X^2 = \frac{(21 - 25)^2}{25} + \frac{(62 - 50)^2}{50} + \frac{(17 - 25)^2}{25} = 6.08$$

- Because $K = 3$, $df = 2$, so

$$C_{0.05} = \{X^2 : X^2 \geq \chi^2_{2,0.95} = 5.99\}$$

- Also,

$$\Pr[\chi^2_2 \geq 6.08] = 1 - 0.952 = 0.048$$

- Reject H_0

χ^2 GOF: Genetics Example Using SAS

```
proc freq order=data;  
  tables genotype / testp=(25 50 25);  
  weight count;
```

The FREQ Procedure

genotype	Frequency	Percent	Test	Cumulative	Cumulative
			Percent	Frequency	Percent
dominant	21	21.00	25.00	21	21.00
heterozygous	62	62.00	50.00	83	83.00
recessive	17	17.00	25.00	100	100.00

Chi-Square Test
for Specified Proportions

Chi-Square 6.0800
DF 2
Pr > ChiSq 0.0478

χ^2 GOF: DBP Example

- Diastolic blood pressure (DBP) was measured on a random sample from a population of interest
- It is hypothesized that DBP is normally distributed

DBP	Frequency
<50	57
[50, 60)	330
[60, 70)	2132
[70, 80)	4584
[80, 90)	4604
[90, 100)	2119
[100, 110)	659
≥ 110	251
Total	14736

χ^2 GOF: DBP Example cont.

- From the sample (before classifying into intervals)

$$\bar{y} = 80.7 \text{ and } s = 12.00$$

- If DBP is normally distributed with $\mu = 80.7$ and $\sigma = 12$, the expected frequency in an interval between a and b is

$$14736 \times [\Phi((b - 80.7)/12) - \Phi((a - 80.7)/12)]$$

- The expected frequency in the <50 group is

$$14736 \Phi((50 - 80.7)/12) = 14736 \Phi[-2.56] = 77.5$$

χ^2 GOF: DBP Example cont.

DBP	Freq	z	$\Phi(z)$	Prob	E
< 50	57	-2.56	0.0053	0.0053	77.5
[50, 60)	330	-1.73	0.0423	0.0370	545.3
[60, 70)	2,132	-0.89	0.1863	0.1440	2,122.3
[70, 80)	4,584	-0.06	0.4767	0.2905	4,280.2
[80, 90)	4,604	0.77	0.7808	0.3041	4,481.0
[90, 100)	2,119	1.61	0.9461	0.1653	2,435.7
[100, 110)	659	2.44	0.9927	0.0466	686.3
≥ 110	251	∞	1	0.0073	107.7
Total	14,736				14,736.0

χ^2 GOF: DBP Example cont.

- H_0 : data are from a normal distribution
- H_A : data are not from a normal distribution
- Rejection region

$$C_\alpha = \{X^2 : X^2 \geq \chi_{K-S-1,1-\alpha}^2\}$$

where S = number of parameters estimated

- For the DBP example, $K = 8$ and $S = 2$, so that

$$C_{0.05} = \{X^2 : X^2 \geq \chi_{5,0.95}^2 = 11.07\}$$

χ^2 GOF: DBP Example cont.

- GOF test statistic

$$X^2 = \sum_{i=1}^K \frac{(O_i - E_i)^2}{E_i} = \frac{(57 - 77.5)^2}{77.5} + \dots = 348.3$$

- Reject H_0
- Read Section 6.6.4 of the text about the distribution of X^2 when parameter estimation is necessary

χ^2 GOF: Mendel Example

- Fisher examined Mendel's experiments (text, Table 6.9)

Experiment	X^2	DF	p
3:1 Ratios	2.14	7	0.95
2:1 Ratios	5.17	8	0.74
Bifactorial	2.81	8	0.95
Gametic ratios	3.67	15	0.999
Trifactorial	15.32	26	0.95
Total	29.11	64	

- If X_1^2, \dots, X_n^2 are independent χ^2 with m_1, \dots, m_n degrees of freedom, then

$$\sum_i X_i^2 \sim \chi_m^2 \quad \text{where} \quad m = \sum_i m_i$$

$$p = \Pr[\chi_{64}^2 > 29.11] = 0.9999474$$

Measurement of Agreement

- Example: Adults were asked to rate their weight as underweight, normal, overweight, or obese; their weight was then measured

Self-report	Measured				Total
	Under	Normal	Over	Obese	
Under	462	178	0	0	640
Normal	72	2868	505	2	3447
Over	0	134	2086	280	2500
Obese	0	0	59	809	868
Total	534	3180	2650	1091	7455

Measure of Agreement: Kappa

- The χ^2 test for independence will reject H_0
- If we want to measure agreement, we might take the proportion on the diagonal:

$$p_a = \frac{462 + 2868 + 2086 + 809}{7455} = 0.835$$

- However, there would be some agreement by chance even if the two classifications were independent

Measure of Agreement: Kappa

- Under independence,

$$E_{11} = (640)(534)/7455 = 45.84$$

$$E_{22} = 1470.35, \quad E_{33} = 888.67, \quad E_{44} = 127.03$$

- Therefore we expect 2531.89 agreements just by chance

$$p_c = \frac{2531.89}{7455} = 0.340$$

Measure of Agreement: Kappa

- Let

p_a = observed proportion of agreement

p_c = expected proportion of agreement

- Kappa statistic

$$\kappa = \frac{p_a - p_c}{1 - p_c}$$

- κ is a chance-adjusted measure of agreement

Measure of Agreement: Kappa

- Note

$$\frac{-p_c}{1 - p_c} \leq \kappa \leq 1$$

$\kappa = 0$ if agreement is totally by chance

$\kappa = 1$ if and only if there is perfect agreement

- There are various categorizations (or guidelines) of values of κ ; the best known is by Landis & Koch (1977). For instance, 0.41-0.60 is classified as “moderate agreement” and 0.61-0.80 as “substantial agreement”.
- My opinion is that the interpretation of κ needs to be context-dependent.

Measure of Agreement: Kappa

- Under $H_0 : \kappa = 0$,

$$\text{Var}(\kappa) = \frac{p_c + p_c^2 - N^{-3} \sum_{i=1}^r (n_{i\cdot}^2 n_{\cdot i} + n_{i\cdot} n_{\cdot i}^2)}{N(1-p_c)^2}$$

- For moderate sample sizes,

$$z = \frac{\kappa}{\sqrt{\text{Var}(\kappa)}} \sim N(0, 1)$$

under H_0

Measure of Agreement: Kappa

- Weight example revisited:

$$\kappa = \frac{0.835 - 0.34}{1 - 0.34} = 0.75$$

which Landis & Koch would classify as indicating substantial agreement

- Compute variance of κ

$$\text{Var}(\kappa) = \frac{0.34 + 0.34^2 - 0.2631}{7455(0.66)^2} = 5.901 \times 10^{-5}$$

- Therefore

$$z = \frac{0.75}{0.0077} = 97.654$$

Kappa: SAS

```
proc freq order=data;  
  tables self*measured/agree nopct norow nocol;  
  test kappa;  
  weight wt;
```

Simple Kappa Coefficient

Kappa	0.7502
ASE	0.0065
95% Lower Conf Limit	0.7374
95% Upper Conf Limit	0.7629

Test of H0: Kappa = 0

ASE under H0	0.0077
Z	97.6540
One-sided Pr > Z	<.0001
Two-sided Pr > Z	<.0001

Kappa: R

- R has kappa statistics in various libraries. The one below is in the vcd library. Note the uppercase K in Kappa()

```
> table<-matrix(c(462,178,0,0,72,2868,505,2,0,134,2086,280,  
+ 0,0,59,809),nrow=4,byrow=T)
```

```
> Kappa(table)  
          value        ASE  
Unweighted 0.7501581 0.006509663  
Weighted   0.8109287 0.013309715
```

```
> confint(Kappa(table))  
  
Kappa          lwr        upr  
Unweighted 0.7373994 0.7629169  
Weighted   0.7848422 0.8370153
```

Kappa: R cont.

- R also has a function `kappa2()` in the library `irr`; but it requires the data in a different format – an $n \times 2$ matrix with each row corresponding to an observation (a pair)

```
> col1<-c(rep('under',640),rep('normal',3447),rep('over',2500),
+   rep('obese',868))

> col2<-c(rep('under',462),rep('normal',178),rep('under',72),
+   rep('normal',2868),rep('over',505),rep('obese',2),
+   rep('normal',134),rep('over',2086),rep('obese',280),
+   rep('over',59),rep('obese',809))

> tab2<-cbind(col1,col2)
```

Kappa: R cont.

```
> rbind(tab2[1:5,],tab2[4081:4090,])
```

	col1	col2
[1,]	"under"	"under"
[2,]	"under"	"under"
[3,]	"under"	"under"
[4,]	"under"	"under"
[5,]	"under"	"under"
[6,]	"normal"	"over"
[7,]	"normal"	"over"
[8,]	"normal"	"over"
[9,]	"normal"	"over"
[10,]	"normal"	"over"
[11,]	"normal"	"obese"
[12,]	"normal"	"obese"
[13,]	"over"	"normal"
[14,]	"over"	"normal"
[15,]	"over"	"normal"

Kappa: R cont.

```
> kappa2(tab2)
```

Cohen's Kappa for 2 Raters (Weights: unweighted)

Subjects = 7455

Raters = 2

Kappa = 0.75

z = 97.7

p-value = 0

```
> agree(tab2)
```

Percentage agreement (Tolerance=0)

Subjects = 7455

Raters = 2

%-agree = 83.5

BIOS 662 Fall 2018

Goodness-of-Fit Tests

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Assessing Fit

- Graphical displays such as qqplot
- Tests
 - χ^2
 - Kolmogorov-Smirnov one-sample (page 279 of the text)
 - Others

Kolmogorov-Smirnov Goodness-of-Fit Test

- Kolmogorov-Smirnov goodness-of-fit test (one sample test)
- We want to test whether our data come from a known and completely specified distribution: $F_0(y)$

Kolmogorov-Smirnov Goodness-of-Fit Test

- The empirical distribution function (EDF) for a given data set is

$$F_n(y) = \begin{cases} 0 & \text{if } y < y_{(1)} \\ k/n & \text{if } y_{(k)} \leq y < y_{(k+1)} \\ 1 & \text{if } y > y_{(n)} \end{cases}$$

Note: The text calls this the *empirical cumulative distribution* (ECD) – Definition 3.9 on page 32

Kolmogorov-Smirnov Goodness-of-Fit Test

- $H_0: Y_1, \dots, Y_n \sim F_0(y)$
- The KS statistic for goodness-of-fit is

$$D = \max_y |F_0(y) - F_n(y)|$$

- Exact and asymptotic distributions of D have been derived, tabulated
- Critical values on the next page are appropriate for continuous $F_0(y)$

Kolmogorov-Smirnov Goodness-of-Fit Test

- Critical values for the KS one sample test

n	$\alpha = 0.05$	$\alpha = 0.01$
10	0.409	0.489
15	0.338	0.404
16	0.327	0.392
17	0.318	0.381
18	0.309	0.371
19	0.301	0.363
20	0.294	0.352
25	0.264	0.317
30	0.242	0.290
35	0.224	0.269
>35	$\frac{1.36}{\zeta}$	$\frac{1.63}{\zeta}$

where $\zeta = (n + \sqrt{n/10})^{1/2}$. Source: Conover, *Practical Nonparametric Statistics*, 1980, page 462.

Kolmogorov-Smirnov Goodness-of-Fit Test

- The KS statistic for goodness-of-fit is

$$D = \max_y |F_0(y) - F_n(y)|$$

- Equivalently

$$D = \max\{D_1, \dots, D_n\}$$

where

$$D_i \equiv \max \left(\frac{i}{n} - x_{(i)}, x_{(i)} - \frac{(i-1)}{n} \right)$$

and

$$x_{(i)} = F_0(y_{(i)})$$

KS GOF Test: Example

- Consider this random sample of size 10:

y_1	0.621
y_2	0.503
y_3	0.203
y_4	0.477
y_5	0.710
y_6	0.581
y_7	0.329
y_8	0.480
y_9	0.554
y_{10}	0.382

KS GOF Test: Example cont.

- It is hypothesized that this sample is from the $U(0, 1)$ distribution

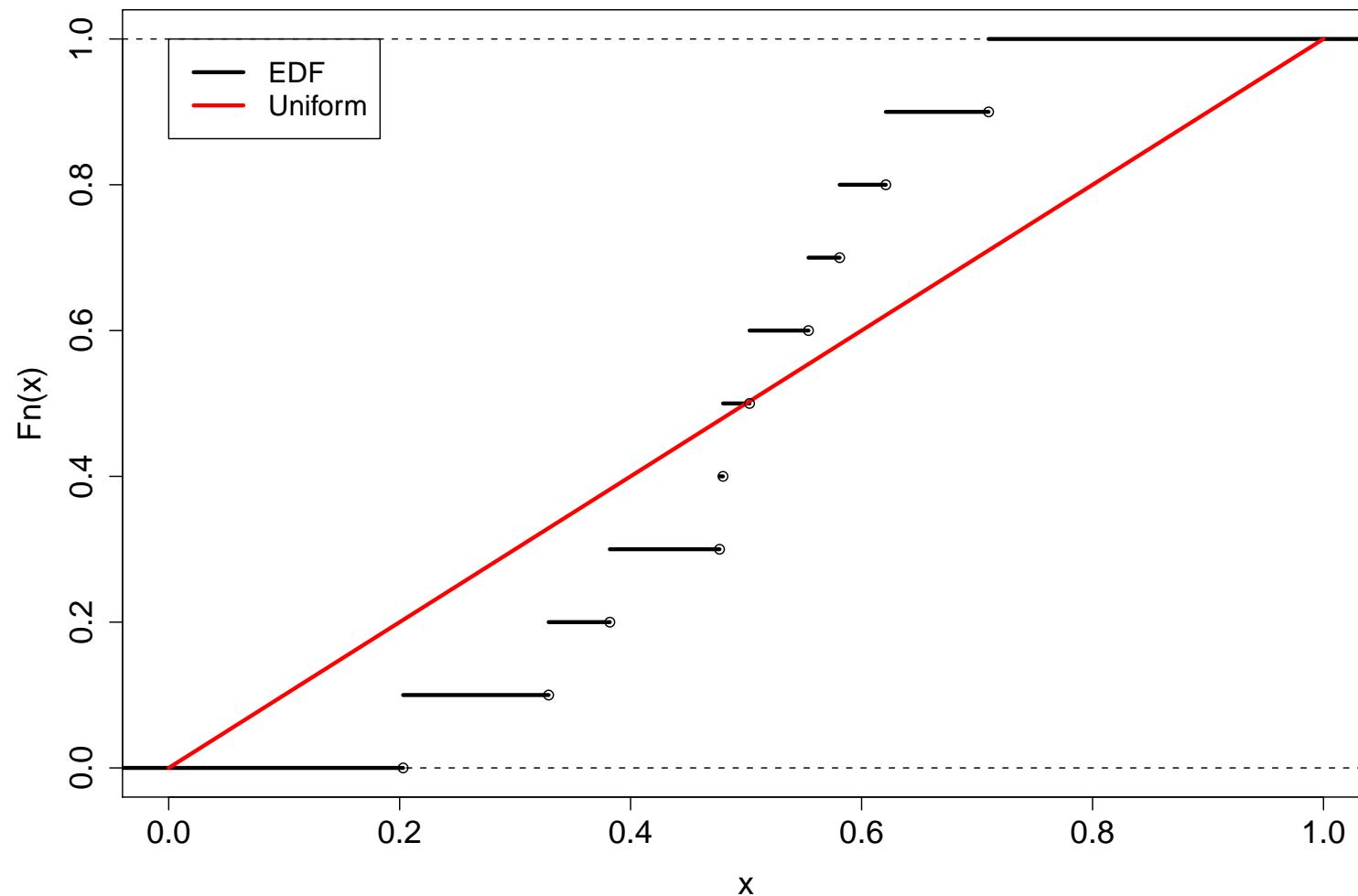
$$F_0(y) = \begin{cases} 0 & \text{if } y < 0 \\ y & \text{if } 0 \leq y \leq 1 \\ 1 & \text{if } 1 < y \end{cases}$$

- $n = 10$
- $C_{0.05} = \{D : D > 0.409\}$
- On the next page we show that $D = 0.290$; thus we do not reject H_0

KS GOF Test: Example cont.

$y_{(i)}$	$F_0(y_{(i)})$	i/n	$(i - 1)/n$	D_i
$y_{(1)}$	0.203	0.1	0.0	0.203
$y_{(2)}$	0.329	0.2	0.1	0.229
$y_{(3)}$	0.382	0.3	0.2	0.182
$y_{(4)}$	0.477	0.4	0.3	0.177
$y_{(5)}$	0.480	0.5	0.4	0.080
$y_{(6)}$	0.503	0.6	0.5	0.097
$y_{(7)}$	0.554	0.7	0.6	0.146
$y_{(8)}$	0.581	0.8	0.7	0.219
$y_{(9)}$	0.621	0.9	0.8	0.279
$y_{(10)}$	0.710	1.0	0.9	0.290

KS GOF Test: Example cont.



Kolmogorov-Smirnov Goodness-of-Fit Test

- The KS test requires that the parameters of $F_0(y)$ are known
- If they are estimated from the data, the distribution of D is not as in the table several pages back
- Critical values for KS statistic for testing normality when μ and σ^2 are estimated are given by Lilliefors (JASA 1967, p. 399)

Lilliefors KS GOF Test

- Critical values for KS test of normality

n	$\alpha = 0.05$	$\alpha = 0.01$
10	0.258	0.294
15	0.220	0.257
16	0.213	0.250
17	0.206	0.245
18	0.200	0.239
19	0.195	0.235
20	0.190	0.231
25	0.173	0.200
30	0.161	0.187
>30	$\frac{0.886}{\sqrt{n}}$	$\frac{1.031}{\sqrt{n}}$

- Source: Conover, *Practical Nonparametric Statistics*, 1980, page 463.

KS GOF: Example

- Consider this random sample of size 10:

y_1	0.621
y_2	0.503
y_3	0.203
y_4	0.477
y_5	1.160
y_6	0.581
y_7	0.329
y_8	0.480
y_9	0.554
y_{10}	0.382

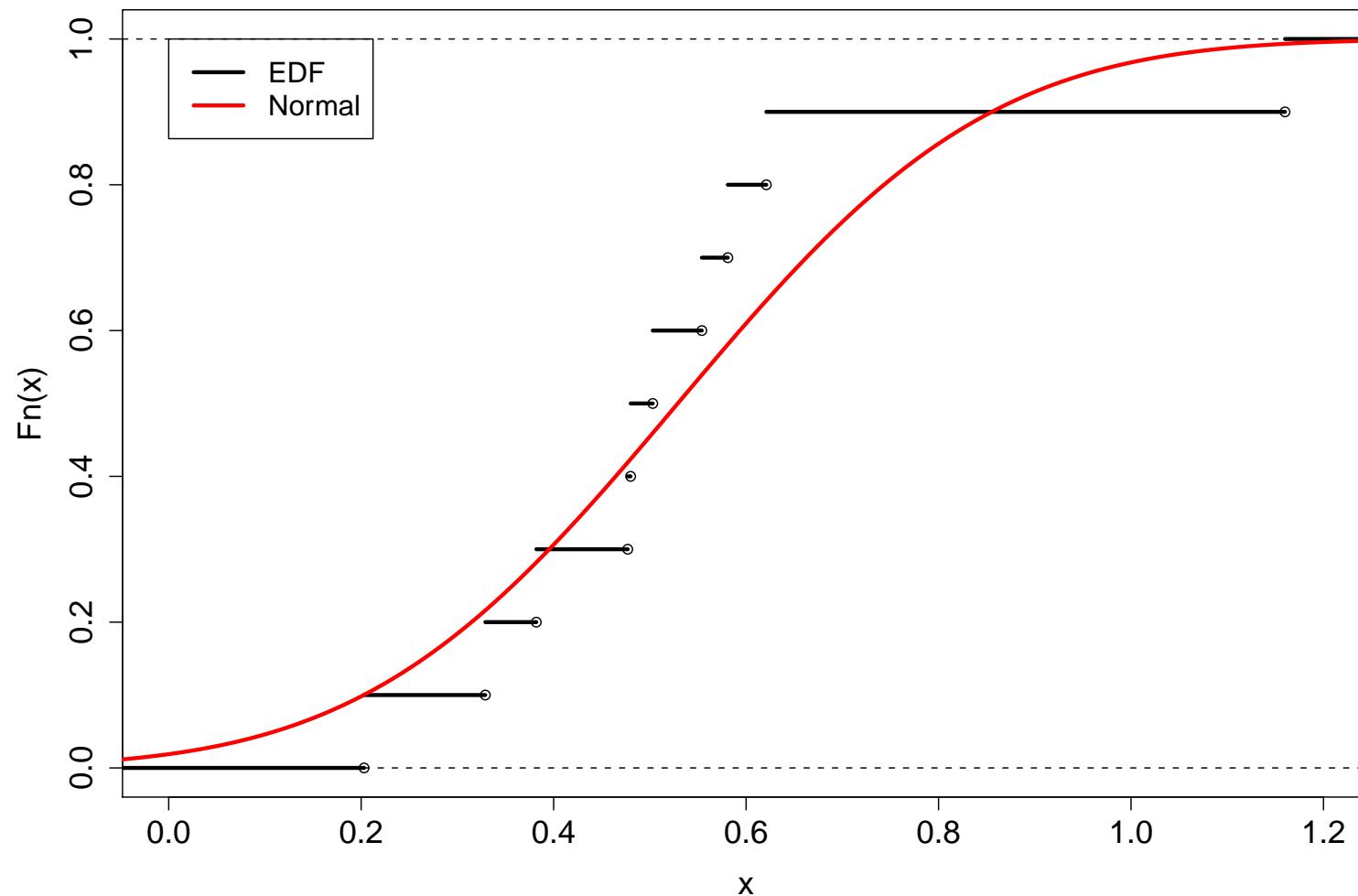
KS GOF: Example cont.

- It is hypothesized that this sample is from a normal distribution
- $\hat{\mu} = \bar{y} = 0.529$ and $\hat{\sigma} = s = 0.2546501$
- $C_{0.05} = \{D : D > 0.258\}$
- For these data $D = 0.259$; $p \approx 0.05$

KS GOF: Example cont.

	$y_{(i)}$	$F_0(y_{(i)})$	i/n	$(i - 1)/n$	D_i
$y_{(1)}$	0.203	0.100	0.1	0.0	0.100
$y_{(2)}$	0.329	0.216	0.2	0.1	0.116
$y_{(3)}$	0.382	0.282	0.3	0.2	0.082
$y_{(4)}$	0.477	0.419	0.4	0.3	0.119
$y_{(5)}$	0.480	0.424	0.5	0.4	0.076
$y_{(6)}$	0.503	0.459	0.6	0.5	0.141
$y_{(7)}$	0.554	0.539	0.7	0.6	0.161
$y_{(8)}$	0.581	0.581	0.8	0.7	0.219
$y_{(9)}$	0.621	0.641	0.9	0.8	0.259
$y_{(10)}$	1.160	0.993	1.0	0.9	0.093

KS GOF: Example cont.



KS GOF: SAS

- SAS: use proc univariate with the option “normal” or the “histogram” statement

```
proc univariate normal;  
var x;
```

Tests for Normality

Test	--Statistic---		-----p Value-----	
Shapiro-Wilk	W	0.835123	Pr < W	0.0386
Kolmogorov-Smirnov	D	0.258945	Pr > D	0.0560
Cramer-von Mises	W-Sq	0.116363	Pr > W-Sq	0.0587
Anderson-Darling	A-Sq	0.710057	Pr > A-Sq	0.0444

KS GOF: SAS

```
proc univariate;  
    histogram x / normal;  
    ** The plot isn't meaningful with so few observations;  
    ** The table has a different heading;
```

The UNIVARIATE Procedure

Fitted Normal Distribution for x

Goodness-of-Fit Tests for Normal Distribution

Test	-----Statistic-----	-----p Value-----
Kolmogorov-Smirnov	D 0.25894505	Pr > D 0.056
Cramer-von Mises	W-Sq 0.11636316	Pr > W-Sq 0.059
Anderson-Darling	A-Sq 0.71005670	Pr > A-Sq 0.044

KS GOF: R

- R function `ks.test()`; however, beware of ties:

```
> set.seed(34621)
> ks.test(rnorm(100000,0,1),"pnorm",0,1)
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: rnorm(1e+05, 0, 1)
D = 0.0032, p-value = 0.2591
alternative hypothesis: two-sided
```

```
> ks.test(rpois(100000,3),"ppois",3)
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: rpois(1e+05, 3)
D = 0.2243, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Warning message:

```
In ks.test(rpois(1e+05, 3), "ppois", 3) :
  cannot compute correct p-values with ties
```

Lilliefors KS GOF: SAS / R

- SAS: automatic
- R: use “nortest” package

```
> x<-c(0.621,0.503,0.203,0.477,1.16,0.581,0.329,0.480,0.554,0.382)
```

```
> ks.test(x,"pnorm",mean(x),sd(x))
```

One-sample Kolmogorov-Smirnov test

```
data: x  
D = 0.2589, p-value = 0.4402  
alternative hypothesis: two-sided
```

```
> # install.packages("nortest")  
> lillie.test(x)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: x  
D = 0.2589, p-value = 0.05602
```

KS vs χ^2 Goodness-of-Fit Tests

- If data are continuous, KS preferred. Why?
 - If sample size small, KS is exact, whereas χ^2 relies on large sample approximation
 - KS test is more powerful than χ^2 in most situations (Conover, *Practical Nonparametric Statistics*, 1980 p. 346)
 - Do not need to bin
- If discrete/categorical, χ^2 preferred

Other Goodness-of-Fit Tests

- Shapiro-Wilk test for normality: see Conover p. 363,
Tables A.17, A.18

$$\frac{\left[\sum_{i=1}^k a_i (X_{(n-i+1)} - X_{(i)}) \right]^2}{s^2}$$

where s^2 is the sample variance and the a_i are given

- Under the null (i.e., normality), numerator and denominator both estimating (up to a constant) σ^2
- R: `shapiro.test()`

Other Goodness-of-Fit Tests

- Class of goodness-of-fit test statistics

$$n \int \{F_n(y) - F_0(y)\}^2 \psi(y) dy$$

- Anderson-Darling $\psi(y) = \{F_0(y)(1 - F_0(y))\}^{-1}$
- Cramer-von Mises $\psi(y) = 1$
- R nortest package: ad.test(), cvm.test()
- SAS: Automatic with “proc univariate normal;”

BIOS 662 Fall 2018

Poisson Random Variables

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

The Poisson Distribution

- Chapter 6.5 of the text
- Two main applications:
 - Modeling counts of discrete events in space or time
 - Approximation to the Binomial distribution for large N and small p

Poisson - Examples

- Number of abnormal cells in a fixed area of a histological slide
- Count of bacteria surviving treatment in a fixed volume of bacterial suspension
- Number of white blood cells in a drop of blood
- Number of new breast cancer cases registered per month by the National Cancer Registry
- Number of live births in Greater London during the month of January

The Poisson Distribution

- Two assumptions required for the Poisson distribution to be an appropriate model:
 - The number of events occurring in one part of the continuum (space, time) should be statistically independent of the number of events occurring in another part of the continuum
 - The expected number of counts in a given part of the continuum should approach zero as its size approaches zero

The Poisson Distribution

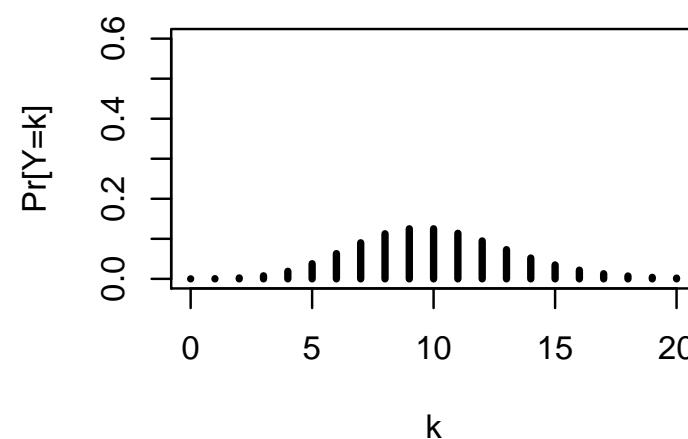
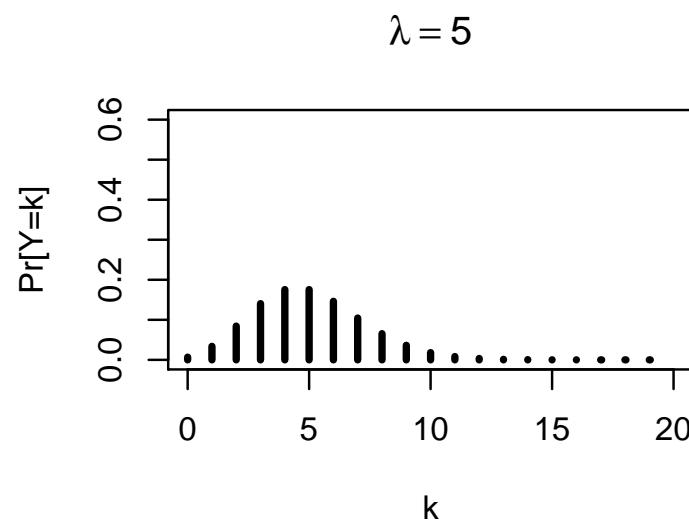
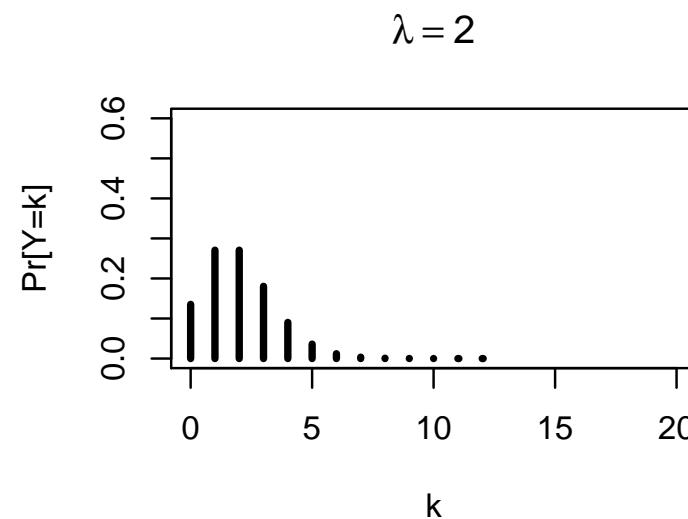
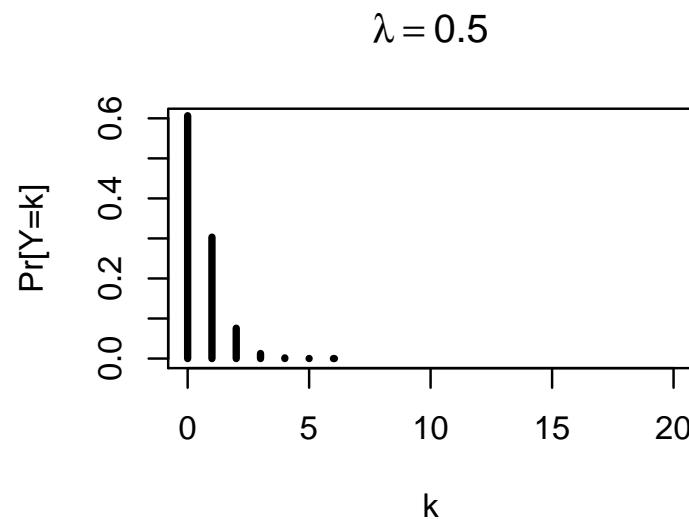
- The Poisson distribution is characterized by one parameter, λ
- If $Y \sim \text{Poisson}(\lambda)$, the probability mass function of Y is

$$\Pr[Y = k] = \frac{e^{-\lambda} \lambda^k}{k!}$$

- $Y \in \{0, 1, 2, \dots\}$
- The parameter λ is both the mean and variance

$$E(Y) = \text{Var}(Y) = \lambda$$

Poisson Probability Mass Function



The Poisson and Binomial Distributions

- Suppose $X \sim \text{Binomial}(N, \pi)$ and $Y \sim \text{Poisson}(\lambda)$ with $\lambda = N\pi$
- Then for N large and π small

$$\Pr[X = k] \approx \Pr[Y = k]$$

i.e.

$$\binom{N}{k} \pi^k (1 - \pi)^{N-k} \approx \frac{e^{-N\pi} (N\pi)^k}{k!}$$

- Rule of thumb: $\pi \leq 0.1$ and $N \geq 20$

The Poisson and Binomial Distributions

- Table 6.6 of the text

Binomial PMF					Poisson
k	$N = 10$	$N = 20$	$N = 40$	$N = 1000$	PMF
	$\pi = 0.20$	$\pi = 0.10$	$\pi = 0.05$	$\pi = 0.002$	$\lambda = 2$
0	0.1074	0.1216	0.1285	0.1351	0.1353
1	0.2684	0.2702	0.2706	0.2707	0.2707
2	0.3020	0.2852	0.2777	0.2709	0.2707
3	0.2013	0.1901	0.1851	0.1806	0.1804
4	0.0881	0.0898	0.0901	0.0902	0.0902
:	:	:	:	:	:

The Poisson and Binomial Distributions

- Sketch of proof: Suppose on average μ events are expected to occur over some fixed time interval
- Divide the interval into N subintervals short enough such that the probability of two events occurring in the same subinterval is very small
- Then the N subintervals approximate a sequence of N Bernoulli trials with success probability μ/N

The Poisson and Binomial Distributions

- Thus the probability of observing exactly x events in the N subintervals is

$$\frac{N(N-1)\cdots(N-x+1)}{x!} \left(\frac{\mu}{N}\right)^x \left(1 - \frac{\mu}{N}\right)^{N-x} \quad (1)$$

- As $N \rightarrow \infty$,

$$N(N-1)\cdots(N-x+1) \approx N^x$$

and

$$\left(1 - \frac{\mu}{N}\right)^{N-x} \approx \left(1 - \frac{\mu}{N}\right)^N \rightarrow e^{-\mu}$$

- Thus (1) is approximately

$$\frac{N^x}{x!} \left(\frac{\mu}{N}\right)^x e^{-\mu} = \frac{e^{-\mu} \mu^x}{x!}$$

Exact Confidence Intervals

- Cf. Note 6.8 of the text (page 195)
- Given y occurrences,

$$\hat{\lambda} = y$$

and an exact $100(1 - \alpha)\%$ CI for λ is

$$\left[\frac{1}{2} \chi^2_{2y; \alpha/2}, \quad \frac{1}{2} \chi^2_{2(y+1); 1-\alpha/2} \right]$$

Normal Approximations

- If $Y \sim \text{Poisson}(\lambda)$ and λ large (say ≥ 100), then

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

- Thus an approximate $100(1 - \alpha)\%$ CI for λ is

$$Y \pm z_{1-\alpha/2} \sqrt{Y}$$

- A better approximation arises from

$$\sqrt{Y} \sim N\left(\sqrt{\lambda}, \frac{1}{4}\right)$$

- For $\lambda \geq 30$, an approximate CI for $\sqrt{\lambda}$ is

$$\sqrt{Y} \pm \frac{z_{1-\alpha/2}}{2}$$

Sum of Poisson Random Variables

- If Y_1, Y_2, \dots, Y_N iid $\text{Poisson}(\lambda)$, then

$$\sum_{i=1}^N Y_i \sim \text{Poisson}(N\lambda)$$

- Estimator for λ

$$\hat{\lambda} = \frac{1}{N} \sum_i Y_i$$

- If (L, U) is a $100(1 - \alpha)\%$ CI for $N\lambda$,
then $(L/N, U/N)$ is a $100(1 - \alpha)\%$ CI for λ .

For example,

$$\hat{\lambda} \pm z_{1-\alpha/2} \sqrt{\sum_i Y_i / N^2}$$

Example 6.20

- Number of bacterial colonies per plate: 72, 69, 63, 59, 59, 53, 51
- The sum is 426 and the mean is 60.86
- Exact 95% CI for 7λ

$$\left[\frac{1}{2} \chi^2_{2 \times 426; 0.025}, \frac{1}{2} \chi^2_{2 \times 427; 0.975} \right] = [386.50, 468.44]$$

- Normal approximations:

$$426 \pm z_{0.975} \times \sqrt{426} = [385.55, 466.45]$$

$$\left[\left(\sqrt{426} - \frac{z_{0.975}}{2} \right)^2, \left(\sqrt{426} + \frac{z_{0.975}}{2} \right)^2 \right] = [386.51, 467.41]$$

- Divide endpoints by $N = 7$ to get 95% CI for λ

Rules of Thumb

- For $\alpha = 0.05$, an approximate 95% CI for $\sqrt{\lambda}$ is

$$\sqrt{Y} \pm \frac{z_{1-\alpha/2}}{2} \approx \sqrt{Y} \pm 1$$

implying an approximate 95% CI for λ is

$$\left[(\sqrt{Y} - 1)^2, (\sqrt{Y} + 1)^2 \right]$$

- If we observe $y = 0$, a two-sided 90% CI for λ is

$$[0, \frac{1}{2}\chi^2_{2; 0.95}] \approx [0, 3.00]$$

- Thus if we observed 0 events out of N trials, the approximate upper bound on a two-sided 90% CI for λ is $3/N$

Homogeneity Test

- Often, observed counts exhibit larger variance than expected under the Poisson model; this is referred to as *over-dispersion*
- This may occur if the assumption of homogeneity of the λ s is not satisfied
- Want to test

$$H_0 : X_1, X_2, \dots, X_k \sim \text{Poisson} (\lambda)$$

Homogeneity Test

- Construct a χ^2 goodness of fit test using the following result
- Suppose $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, 2, \dots, k$
- Then the conditional distribution of (X_1, \dots, X_k) given $\sum_i X_i = N$ is multinomial with cell probabilities

$$\frac{\lambda_i}{\lambda_1 + \dots + \lambda_k} \text{ for } i = 1, 2, \dots, k$$

Homogeneity Test

- Under H_0 ,

$$H_0 : X_1, X_2, \dots, X_k \sim \text{Poisson } (\lambda)$$

the test statistic

$$T = \frac{\sum_i (X_i - \bar{X})^2}{\bar{X}} \sim \chi_{k-1}^2$$

- Equivalent form

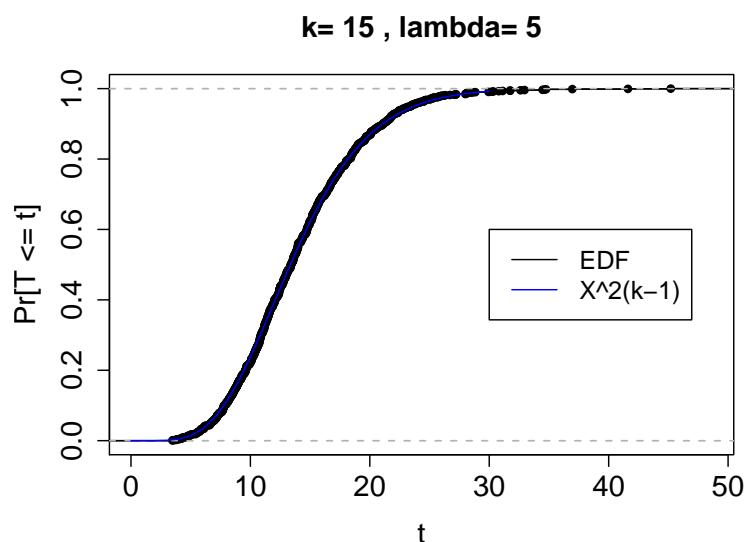
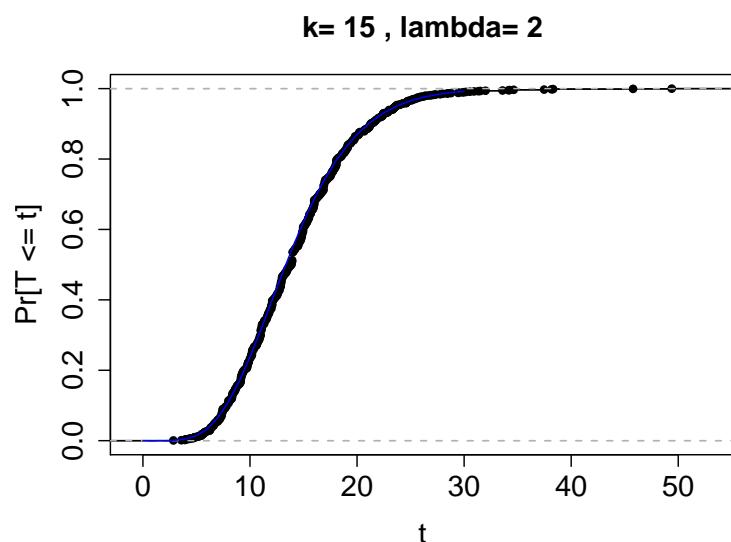
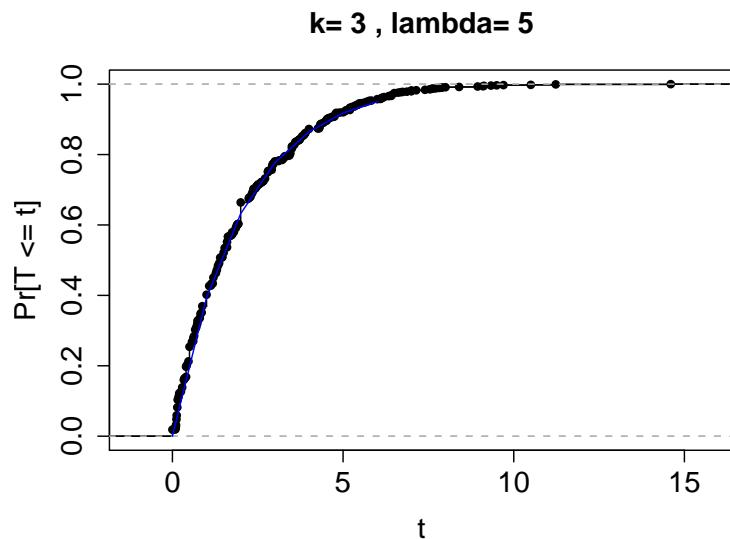
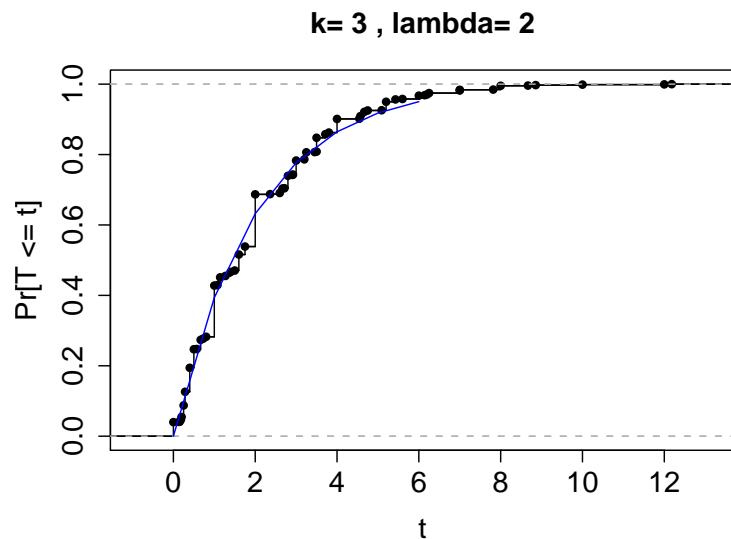
$$T = \frac{(k-1)s^2}{\bar{X}}$$

- *Poisson homogeneity/heterogeneity/dispersion test*

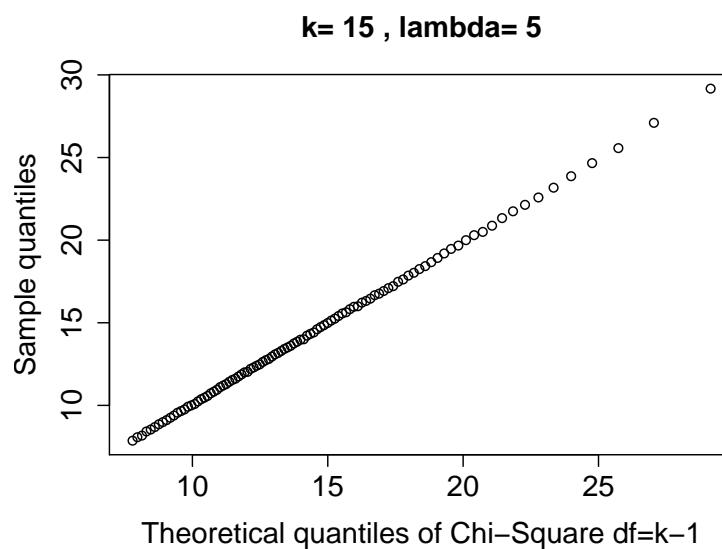
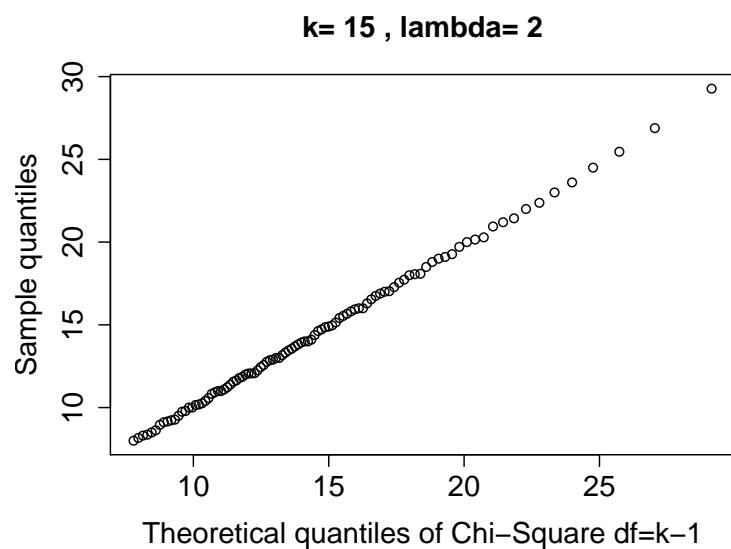
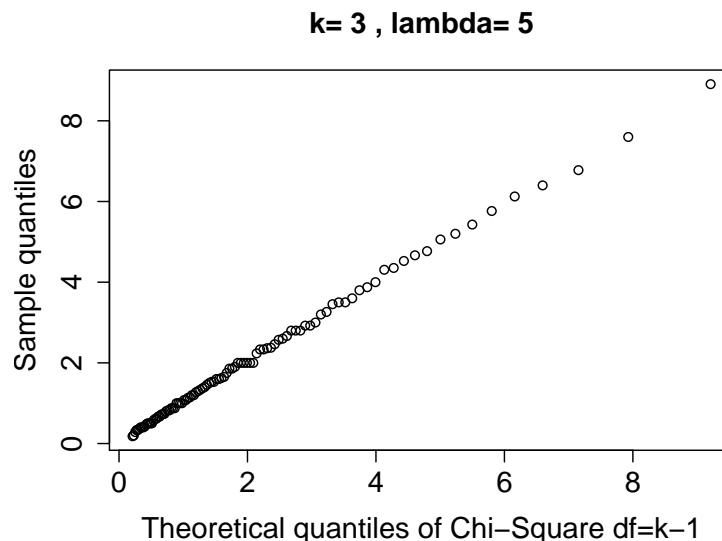
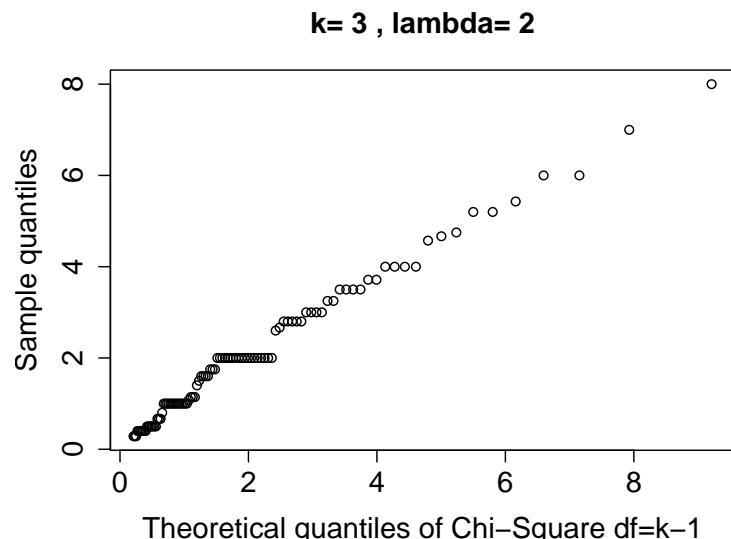
Homogeneity Test

- The χ^2 approximation improves as λ, k get large
- Recommendation (Armitage and Berry 1987)
 1. $\bar{X} \geq 5$, or
 2. $\bar{X} \geq 2$ and $k > 15$

Homogeneity Test: Simulation Study



Homogeneity Test: Simulation Study



Homogeneity Test

- Example 6.20

$$k = 7, \bar{X} = 60.86, s_X = 7.7552$$

implying

$$T = \frac{6 \times (7.7552)^2}{60.86} = 5.93$$

- Because $\Pr[\chi_6^2 > 5.93] = 0.43$, fail to reject H_0

Negative-Binomial Distribution

- Over-dispersion may be due to heterogeneity of λ s
- That is, λ is no longer a constant, but a random variable
- If λ follows a gamma distribution, then the counts follow a negative binomial distribution
- This allows for the variance to be proportional to the mean

BIOS 662 Fall 2018

Linear Regression, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

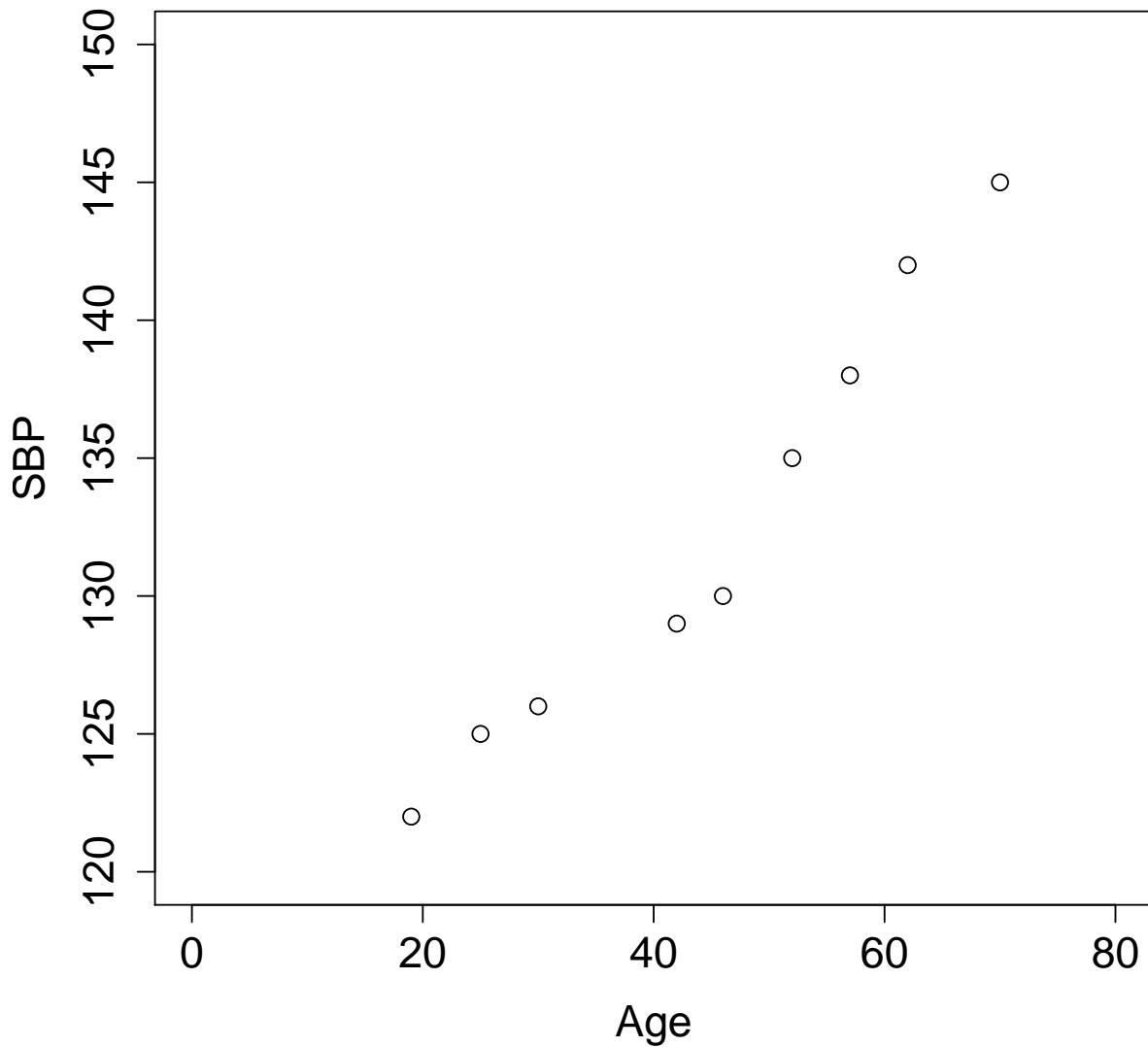
Outline

- Introduction: Assumptions, least-squares estimation
- Confidence intervals and hypothesis testing for regression coefficients
- Confidence interval for mean
- Prediction intervals
- r^2

Example: Systolic Blood Pressure and Age

Obs.	Age	SBP
1	19	122
2	25	125
3	30	126
4	42	129
5	46	130
6	52	135
7	57	138
8	62	142
9	70	145

Example: SBP and Age cont.



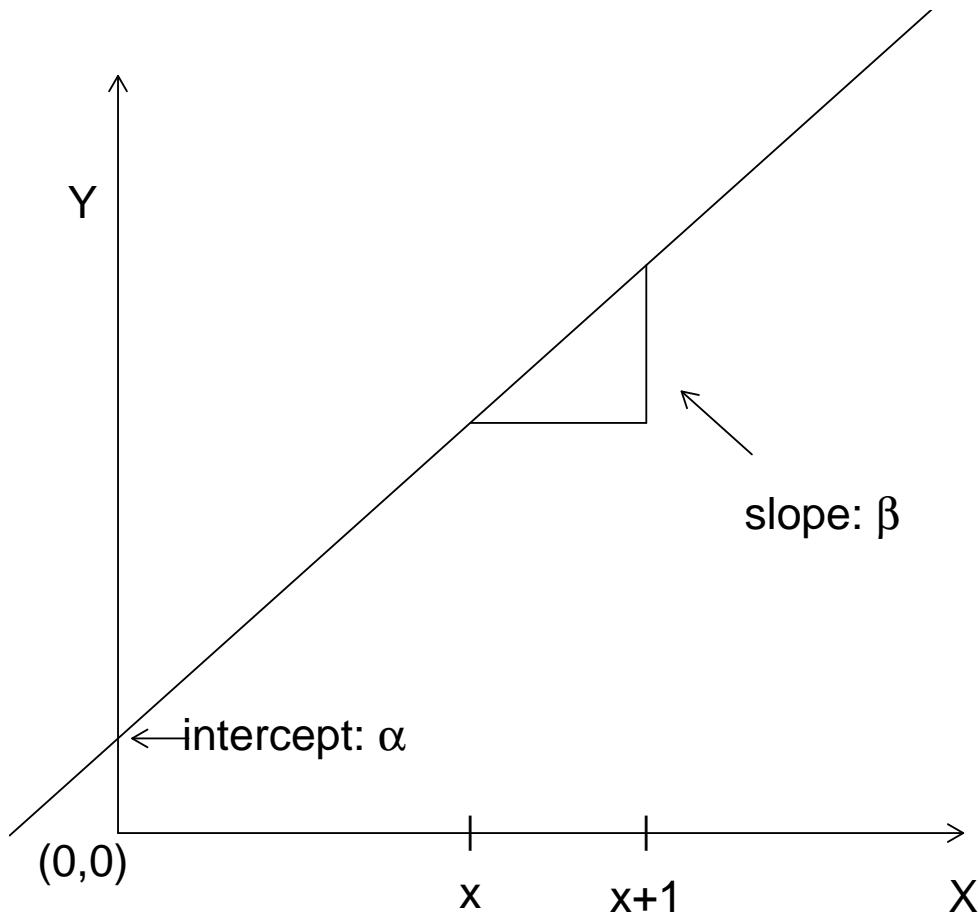
Simple Linear Model

- Line

$$Y = \alpha + \beta X$$

- α = intercept; value of Y when $X = 0$
- β = slope; change in Y when X increases by 1 unit
- Y dependent variable; response variable
- X independent variable; predictor; covariate

Simple Linear Model



Simple Linear Model with Error

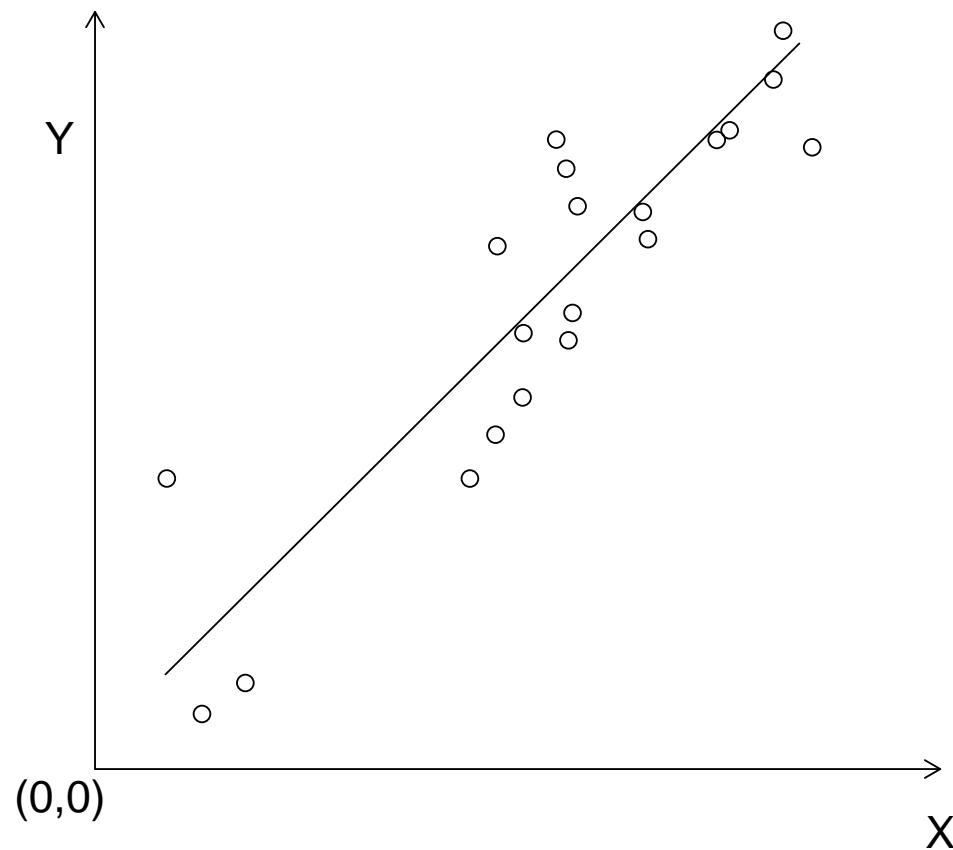
- Linear regression

$$Y = \alpha + \beta X + \epsilon$$

$$\epsilon = Y - \alpha - \beta X$$

- ϵ is the vertical distance from Y to the line defined by $\alpha + \beta X$

Simple Linear Model with Error



Model Assumptions

- Data are $(Y_i, X_i); i = 1, 2, \dots, N$
- Assumptions:
 1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
 2. X s are fixed constants
 3. ϵ_i iid $N(0, \sigma^2)$

Least Squares Estimation

- Least squares estimators are values of α and β that minimize

$$\sum_{i=1}^N \epsilon_i^2 = \sum_{i=1}^N (Y_i - \alpha - \beta X_i)^2$$

- Set partial derivatives equal to 0, solve for α and β
- Can also derive these estimators via maximum likelihood

Least Squares Estimation

- For α :

$$\begin{aligned}\frac{\partial \sum_i \epsilon_i^2}{\partial \alpha} &= -2 \sum_i (Y_i - \alpha - \beta X_i) \\ &= -2N\bar{Y} + 2N\alpha + 2N\beta\bar{X}\end{aligned}$$

- For β :

$$\begin{aligned}\frac{\partial \sum_i \epsilon_i^2}{\partial \beta} &= -2 \sum_i (Y_i - \alpha - \beta X_i) X_i \\ &= -2 \sum_i X_i Y_i + 2\alpha \sum_i X_i + 2\beta \sum_i X_i^2\end{aligned}$$

Least Squares Estimation

- Two equations with two unknowns

$$- 2N\bar{Y} + 2N\alpha + 2N\beta\bar{X} = 0 \quad (1)$$

$$- 2 \sum_i X_i Y_i + 2\alpha \sum_i X_i + 2\beta \sum_i X_i^2 = 0 \quad (2)$$

- From (1)

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

Least Squares Estimation

- Substituting into (2)

$$-\sum_i X_i Y_i + (\bar{Y} - \hat{\beta} \bar{X}) \sum_i X_i + \hat{\beta} \sum_i X_i^2 = 0,$$

implying

$$\hat{\beta} \left(\sum_i X_i^2 - N \bar{X}^2 \right) = \sum_i X_i Y_i - N \bar{X} \bar{Y}.$$

- Therefore

$$\hat{\beta} = \frac{\sum_i X_i Y_i - N \bar{X} \bar{Y}}{\sum_i X_i^2 - N \bar{X}^2}$$

Least Squares Estimation

- Equivalent form:

$$\hat{\beta} = \frac{\sum_i X_i Y_i - N \bar{X} \bar{Y}}{\sum_i X_i^2 - N \bar{X}^2} = \frac{[XY]}{[X^2]}$$

where

$$[XY] = \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

$$[X^2] = \sum_i (X_i - \bar{X})^2$$

- Note that if $X_i = Y_i$ for all i , then $\hat{\beta} = 1$ as one would expect
- Also, if $Y_i = \bar{Y}$ for all i , then $\hat{\beta} = 0$

Least Squares Estimation

- Predicted response (also known as *fitted values*)

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

- Residual

$$r_i = Y_i - \hat{Y}_i$$

- Estimate variance by mean square error (MSE)

$$\begin{aligned}\hat{\sigma}^2 &= s_{y \cdot x}^2 = \frac{1}{N-2} \sum_i (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{N-2} \sum_i r_i^2\end{aligned}$$

Example: SBP and Age

$$\bar{Y} = 132.4; \quad \bar{X} = 44.8$$

$$\sum_i X_i Y_i = 54461; \quad \sum_i X_i^2 = 20463$$

$$\hat{\beta} = \frac{54461 - 9(132.4)(44.8)}{20463 - 9(44.8)^2} = 0.45$$

$$\hat{\alpha} = 132.4 - 0.45(44.8) = 112.3$$

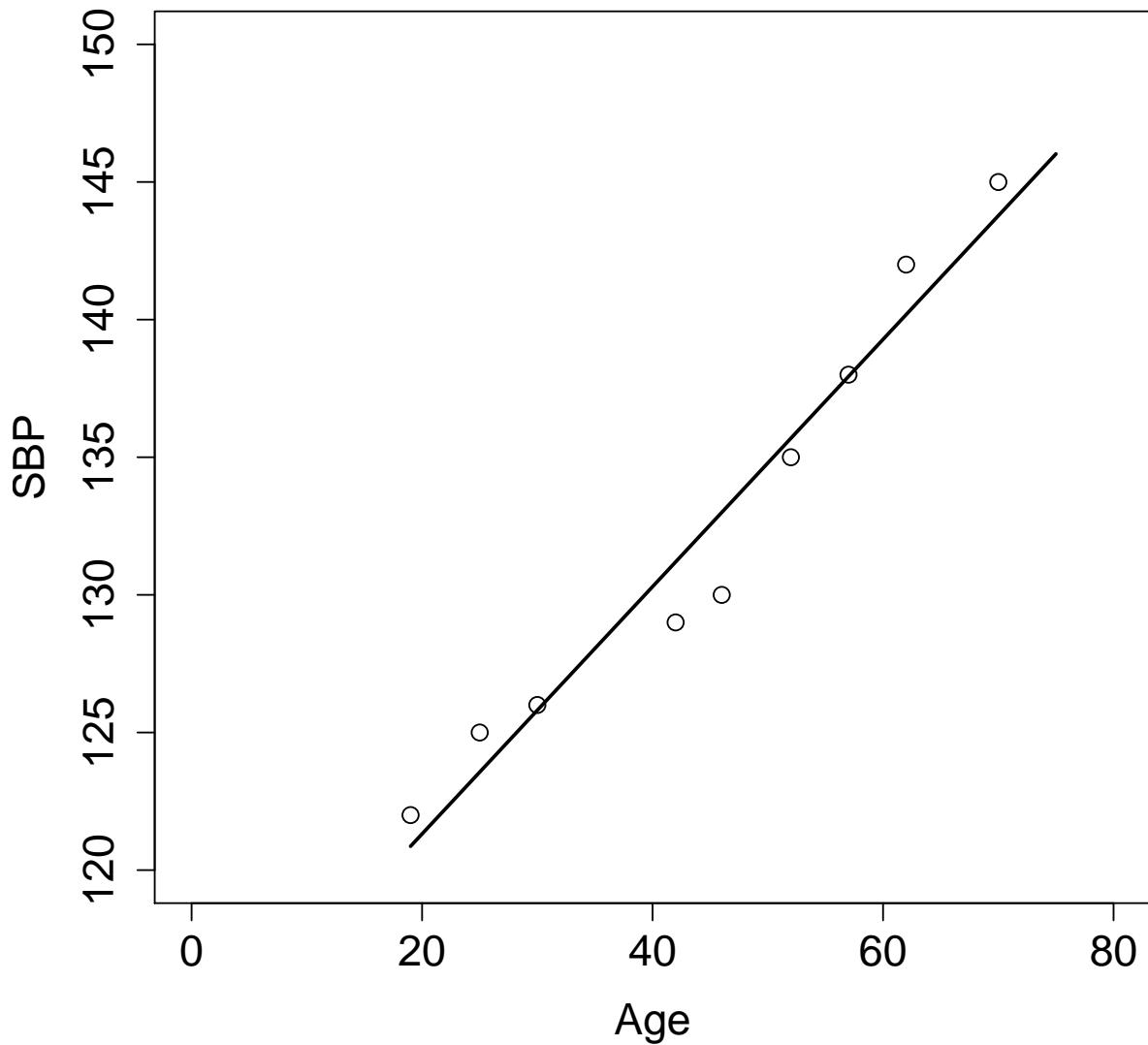
$$\hat{Y}_i = 112.3 + 0.45X_i$$

$$s_{y \cdot x}^2 = 3.21$$

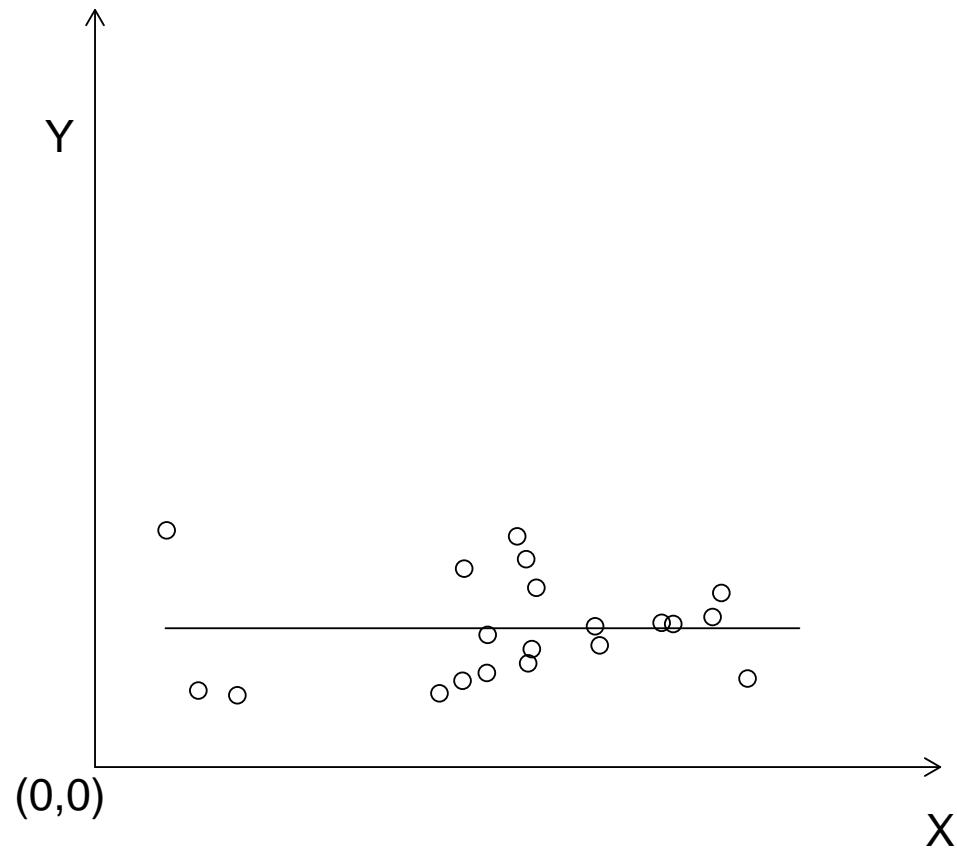
Example: Interpretation

- $\hat{\beta} = 0.45 \Rightarrow$ expected SBP increases 0.45 (mmHg) for each one year increase in age
- $\hat{\alpha} = 112.3 \Rightarrow ?$ Beware extrapolation (see section 9.4.3 of the text)

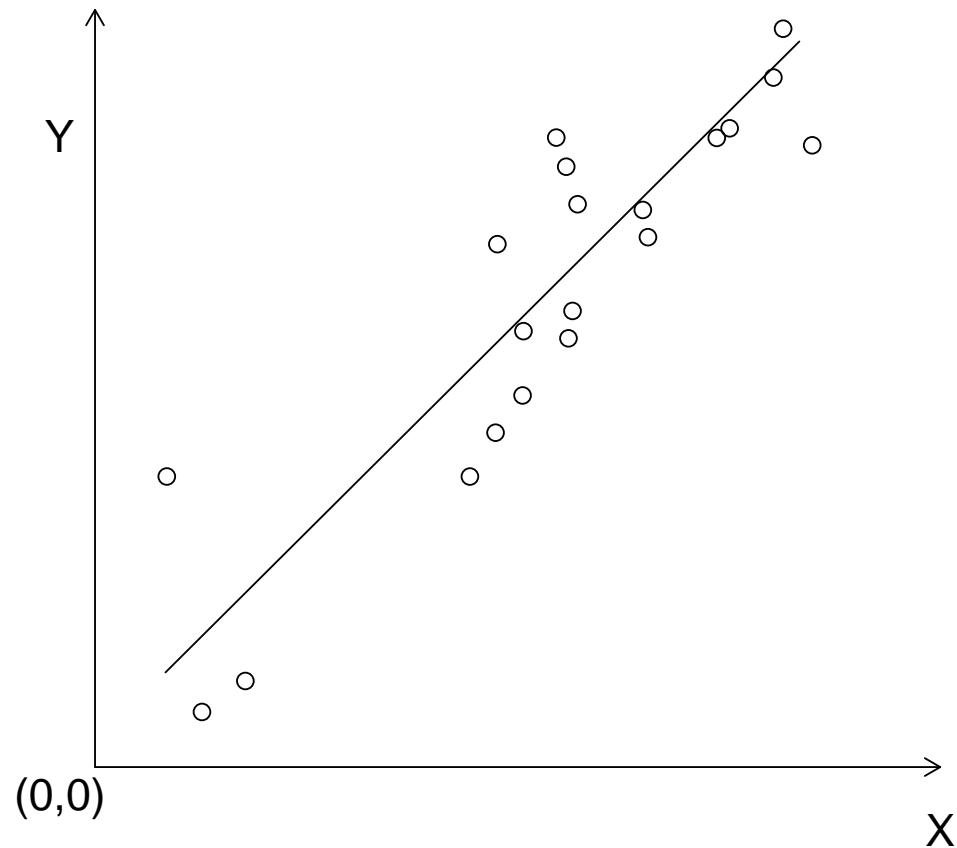
Example: SBP and Age cont.



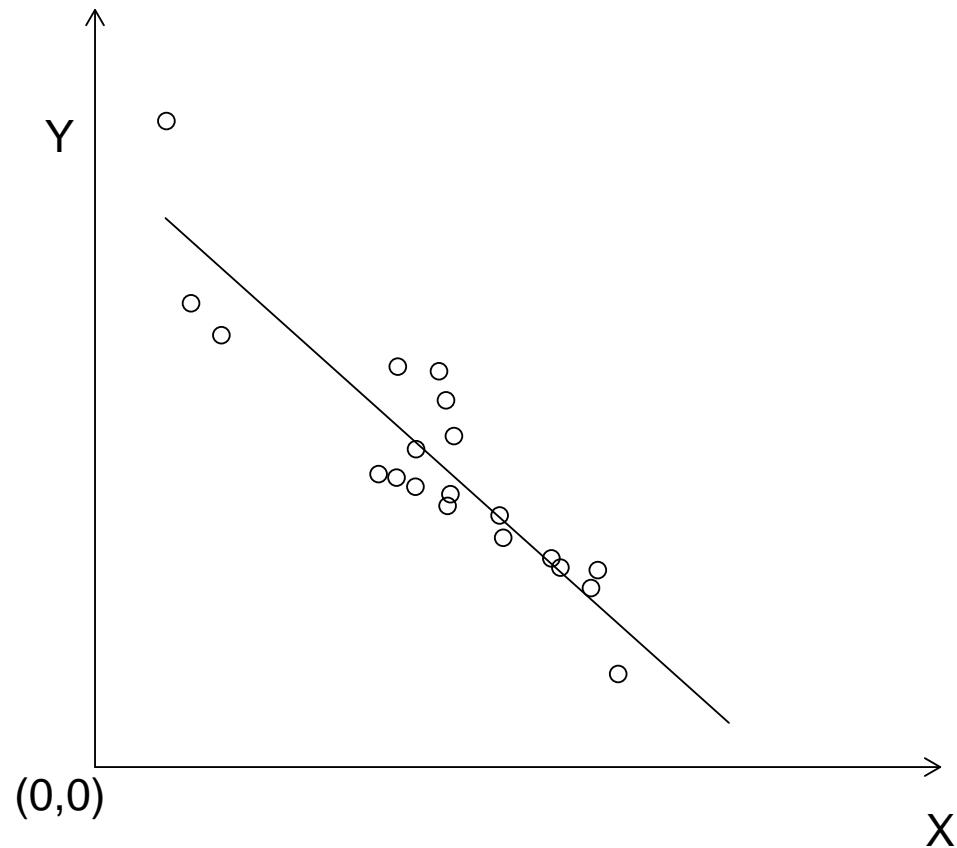
$$\hat{\beta} = 0$$



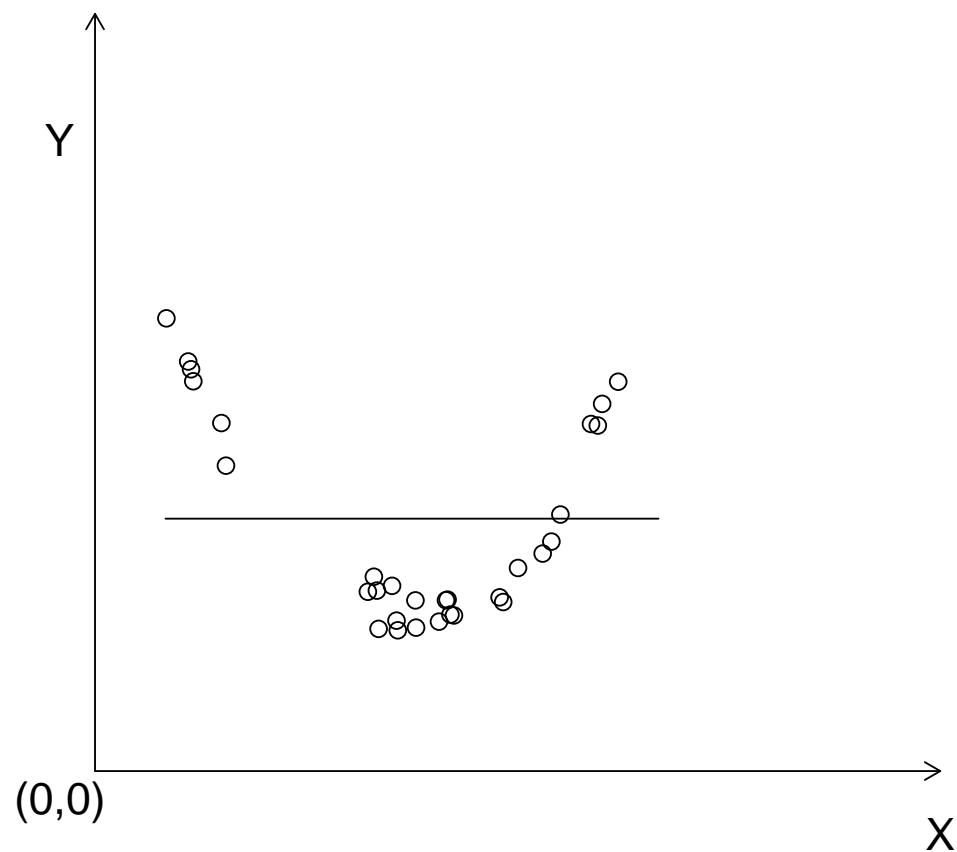
$$\hat{\beta} > 0$$



$$\hat{\beta} < 0$$



$$\hat{\beta} = 0$$



Confidence Intervals and Hypotheses Tests

- Can write

$$\hat{\beta} = \sum c_i Y_i$$

where

$$c_i = \frac{X_i - \bar{X}}{\sum_j (X_j - \bar{X})^2}$$

- Under the model,

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2)$$

- Thus

$$\hat{\beta} \sim N \left(\sum_i c_i (\alpha + \beta X_i), \sigma^2 \sum_i c_i^2 \right)$$

Confidence Intervals and Hypotheses Tests

- Equivalently

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i(X_i - \bar{X})^2}\right)$$

- $100(1 - \alpha)\%$ CI for β

$$\hat{\beta} \pm z_{1-\alpha/2} \sqrt{\frac{\sigma^2}{\sum_i(X_i - \bar{X})^2}}$$

- Test for $H_0 : \beta = \beta_0$

$$z = \frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2 / \sum_i(X_i - \bar{X})^2}}$$

Confidence Intervals and Hypotheses Tests

- If σ^2 is unknown, use $s_{y \cdot x}^2$ and t_{N-2}

- $100(1 - \alpha)\%$ CI for β

$$\hat{\beta} \pm t_{N-2, 1-\alpha/2} \sqrt{s_{y \cdot x}^2 / \sum_i (X_i - \bar{X})^2}$$

- Test for $H_0 : \beta = \beta_0$

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{s_{y \cdot x}^2 / \sum_i (X_i - \bar{X})^2}}$$

Confidence Intervals and Hypotheses Tests: SBP

- For the SBP example, $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$

$$C_{0.05} = \{t : |t| > t_{7,0.975} = 2.365\}$$

- Observed test statistic implies reject H_0

$$t = \frac{0.449 - 0}{\sqrt{3.21/2417.56}} = 12.32$$

- 95% CI

$$0.449 \pm 2.365\sqrt{3.21/2417.56} = (0.363, 0.535)$$

Confidence Intervals and Hypotheses Tests

- It can be shown that \bar{Y} and $\hat{\beta}$ are independent
- Therefore

$$\hat{\alpha} \sim N \left(\alpha, \sigma^2 \left(\frac{1}{N} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2} \right) \right)$$

- $H_0 : \alpha = \alpha_0$

$$t = \frac{\hat{\alpha} - \alpha_0}{s_{y \cdot x} \sqrt{\frac{1}{N} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2}}} \sim t_{N-2}$$

SBP Example in R

```
> fit <- lm(sbp~age)
> summary(fit)
```

Call:

```
lm(formula = sbp ~ age)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.9934	-0.6884	0.1933	1.2265	1.8199

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	112.33169	1.73773	64.64	5.57e-11 ***
age	0.44917	0.03644	12.32	5.31e-06 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.792 on 7 degrees of freedom

Multiple R-Squared: 0.9559, Adjusted R-squared: 0.9497

F-statistic: 151.9 on 1 and 7 DF, p-value: 5.313e-06

SBP Example in SAS

```
proc reg;  
  model sbp=age;
```

The REG Procedure

Model: MODEL1

Dependent Variable: sbp

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	487.74667	487.74667	151.91	<.0001
Error	7	22.47555	3.21079		
Corrected Total	8	510.22222			
Root MSE	1.79187	R-Square	0.9559		
Dependent Mean	132.44444	Adj R-Sq	0.9497		
Coeff Var	1.35292				

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	112.33169	1.73773	64.64	<.0001
age	1	0.44917	0.03644	12.33	<.0001

Confidence Interval for $E(Y|X = x)$

- Goal: CI for the mean of Y given $X = x$
- Let $\mu_x = E(Y|X = x)$
- Estimator for μ_x :

$$\hat{\mu}_x = \hat{\alpha} + \hat{\beta}x$$

$$= \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}x$$

$$= \bar{Y} + \hat{\beta}(x - \bar{X})$$

- $E(\hat{\mu}_x) = \mu_x$

Confidence Interval for $E(Y|X = x)$

- Recall that \bar{Y} and $\hat{\beta}$ are independent normally distributed random variables
- Thus $\hat{\mu}_x$ is normally distributed and

$$\text{Var}(\hat{\mu}_x) = \text{Var}(\bar{Y}) + (x - \bar{X})^2 \text{Var}(\hat{\beta})$$

$$= \frac{\sigma^2}{N} + \frac{\sigma^2(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}$$

$$= \sigma^2 \left[\frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right]$$

Confidence Interval for $E(Y|X = x)$

- Therefore, a $100(1 - \alpha)\%$ CI for μ_x is

$$\hat{\mu}_x \pm t_{N-2, 1-\alpha/2} \sqrt{s_{y|x}^2 \left\{ \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right\}}$$

- Note that $\text{Var}(\hat{\mu}_x)$ is a function of $x - \bar{X}$
- So, the further x is from \bar{X} , the wider the CI will be
- Design considerations: Note 9.3 in the text

Example: SBP and Age

- Suppose we want a 95% CI for the mean SBP when age = 40

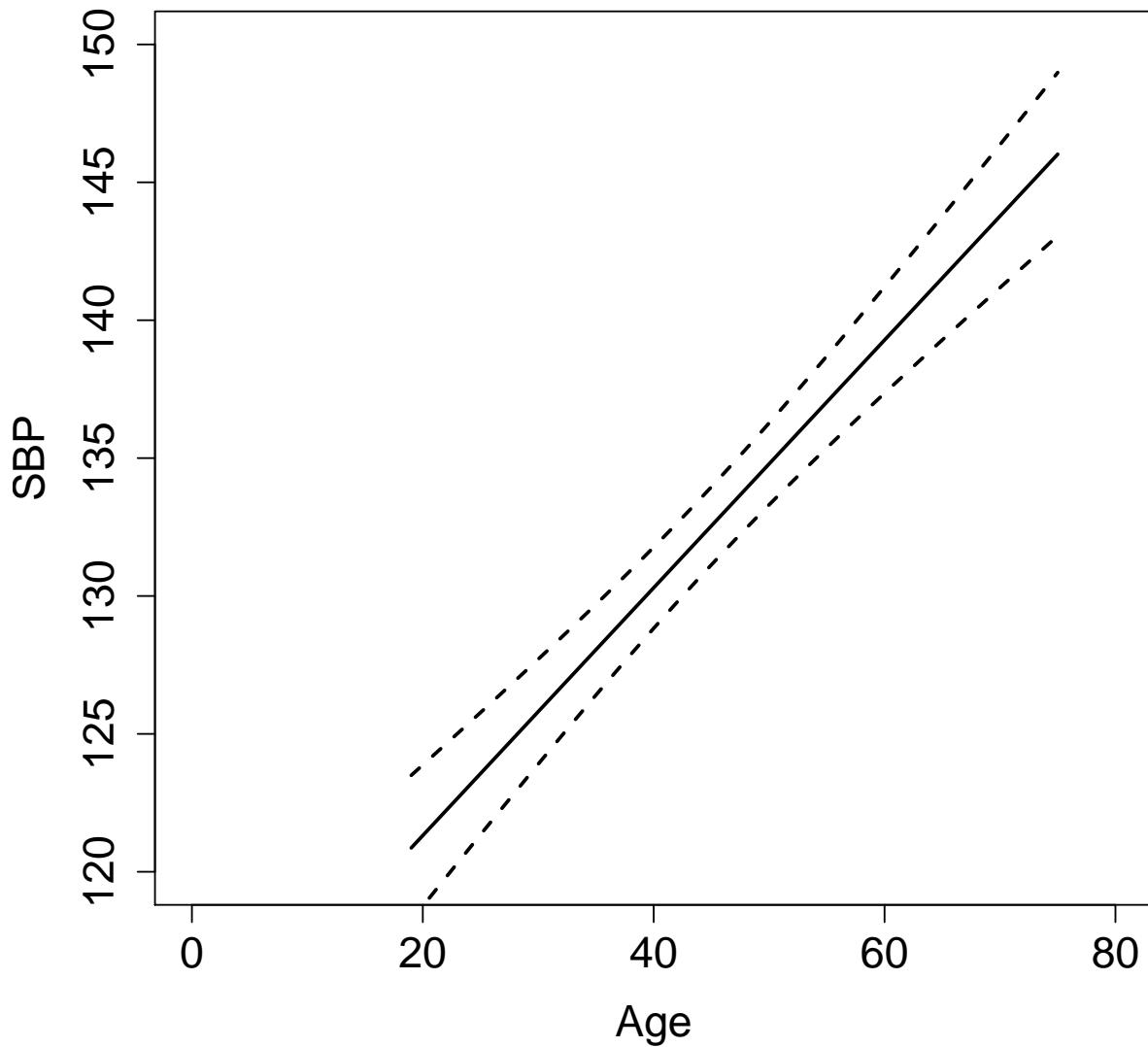
$$\hat{\mu}_{40} = 112.3 + 0.45(40) = 130.3$$

- Confidence interval:

$$130.3 \pm 2.365(1.79) \sqrt{\frac{1}{9} + \frac{(40 - 44.8)^2}{2417.59}}$$

$$(128.8, 131.8)$$

Example: SBP and Age cont.



Confidence Interval for $E(Y|X = x)$

- These “bands” should be interpreted in a pointwise fashion only
- The text’s usage of the term “bands” is non-standard (p. 303-4)
- Usual interpretation of *confidence band*: covers the entire regression line with $100(1 - \alpha)\%$ confidence
- Cf. Section 2.6 of *Applied Linear Statistical Models*, Neter et al., 4th edition, 1996

Prediction

- Suppose we want a prediction interval (PI) for a new or future observation, given $X = x$

$$\hat{Y}_x = \hat{\alpha} + \hat{\beta}x$$

- Note: Y_x is a random variable, so we consider the random variable $Y_x - \hat{Y}_x$

$$E(Y_x - \hat{Y}_x) = \alpha + \beta x - (\hat{\alpha} + \hat{\beta}x) = 0$$

$$\text{Var}(Y_x - \hat{Y}_x) = \text{Var}(Y_x) + \text{Var}(\hat{Y}_x) - 2\text{Cov}(Y_x, \hat{Y}_x)$$

Prediction

- Because Y_x is not part of the sample, Y_x and \hat{Y}_x are independent
- Therefore

$$\begin{aligned}\text{Var}(Y_x - \hat{Y}_x) &= \sigma^2 + \sigma^2 \left(\frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right) \\ &= \sigma^2 \left(1 + \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)\end{aligned}$$

Prediction

- Because ϵ is normally distributed, it follows that

$$Y_x - \hat{Y}_x \sim N \left(0, \sigma^2 \left(1 + \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right) \right)$$

- If σ^2 is not known,

$$\frac{Y_x - \hat{Y}_x}{s_{y \cdot x} \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum(X_i - \bar{X})^2}}} \sim t_{N-2}$$

Prediction

- $100(1 - \alpha)\%$ prediction interval for a new or future observation at $X = x$

$$\hat{Y}_x \pm t_{N-2,1-\alpha/2} s_{y \cdot x} \sqrt{1 + \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum(X_i - \bar{X})^2}}$$

- Cf. Section 5-10 of *Applied Regression Analysis and Multivariable Methods*, Kleinbaum et al., 3rd edition, 1998

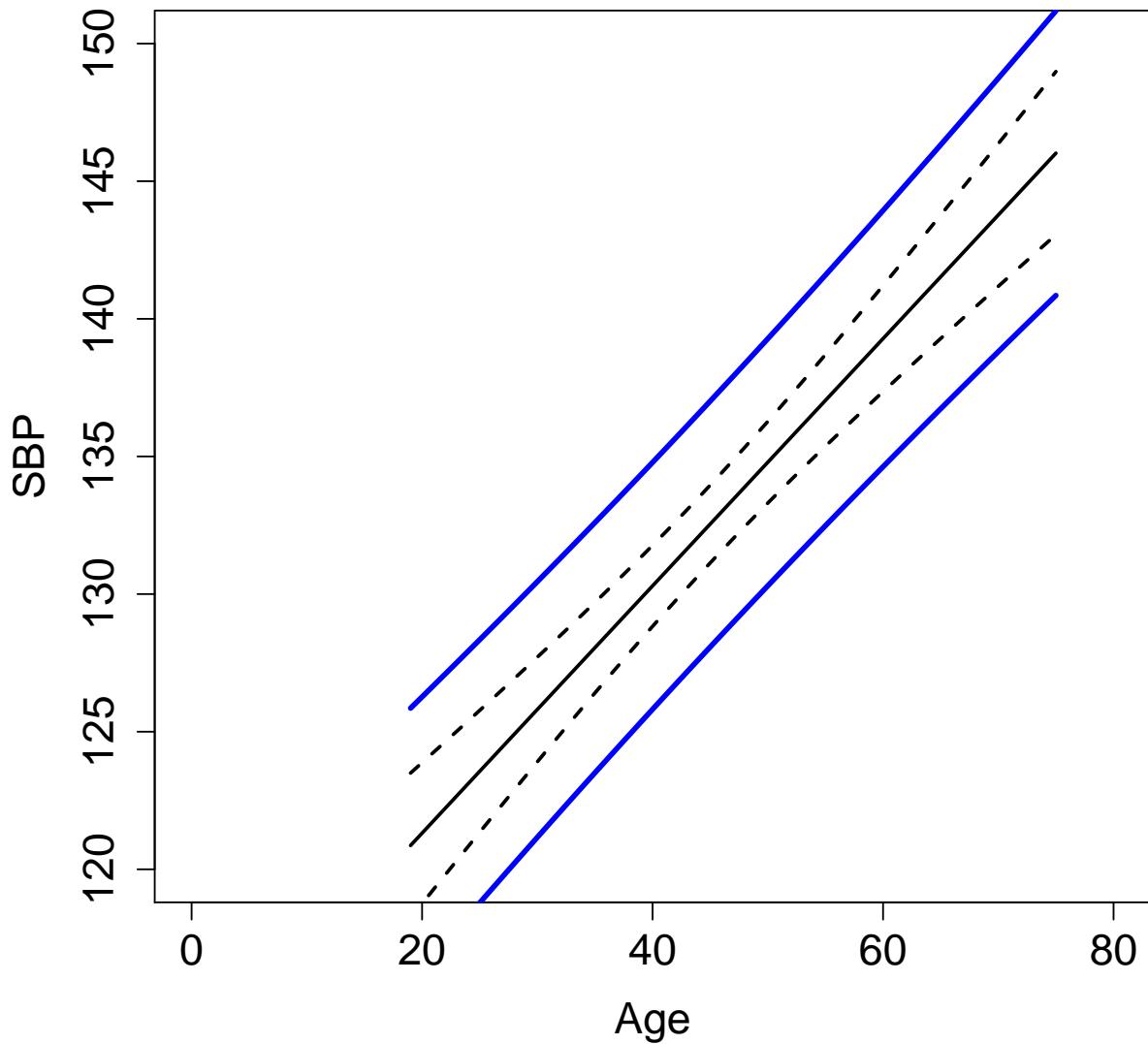
Prediction

- Suppose we want a 95% prediction interval for an individual who is 40 years old
- Point estimate: $\hat{Y}_{40} = 130.3$
- Prediction interval:

$$130.3 \pm 2.365(1.79) \sqrt{1 + \frac{1}{9} + \frac{(40 - 44.8)^2}{2417.59}}$$

$$(125.8, 134.8)$$

Example: SBP vs Age



SBP Example in R

```
> fit <- lm(sbp~age)

> predict(fit,data.frame(age=40),interval="confidence")
      fit      lwr      upr
1 130.2984 128.8273 131.7696

> predict(fit,data.frame(age=40),interval="prediction")
      fit      lwr      upr
1 130.2984 125.8132 134.7836
```

SBP Example in SAS

- In the input dataset add an observation with age = 40 and missing SBP

```
proc reg;  
  model sbp=age;  
  output out=ci lcl=LCL lclm=LCLM p=P uclm=UCLM ucl=UCL;  
  
proc print data=ci;
```

Obs	id	age	sbp	P	LCLM	UCLM	LCL	UCL
1	1	19	122	120.866	118.234	123.498	115.878	125.854
2	2	25	125	123.561	121.347	125.774	118.780	128.341
3	3	30	126	125.807	123.905	127.708	121.162	130.451
4	4	42	129	131.197	129.764	132.629	126.724	135.669
5	5	46	130	132.993	131.577	134.410	128.526	137.461
6	6	52	135	135.688	134.145	137.232	131.179	140.198
7	7	57	138	137.934	136.172	139.696	133.345	142.523
8	8	62	142	140.180	138.131	142.229	135.474	144.887
9	9	70	145	143.773	141.181	146.366	138.806	148.741
10	10	40	.	130.298	128.827	131.770	125.813	134.784

Sum of Squares Decomposition

- We can decompose the total sum of squares

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{Y}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{Y}_i)^2$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

- Total sample variance of the Y s:

$$s_y^2 = \frac{\text{SST}}{N - 1} = \frac{\sum_i (Y_i - \bar{Y})^2}{N - 1}$$

Unadjusted r^2

- The unadjusted r^2 is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- r^2 is called the *coefficient of determination*
- Proportion of total variation attributable to regression
- SBP example:

$$r^2 = \frac{487.75}{510.22} = 0.9559$$

Adjusted r^2

- Note that the sample variance of the Y s is $s_y^2 = 63.78$ while $s_{y \cdot x}^2 = 3.21$
- Thus X “explains” the proportion

$$\frac{63.78 - 3.21}{63.78} = 0.9497$$

of the variance of Y

- This quantity is called the *adjusted r^2*

$$r_a^2 = \frac{s_y^2 - s_{y \cdot x}^2}{s_y^2} = 1 - \frac{s_{y \cdot x}^2}{s_y^2} = 1 - \frac{\text{SSE}/(N - 2)}{\text{SST}/(N - 1)}$$

Adjusted and Unadjusted r^2

- Note that

$$r_a^2 = 1 - \frac{\text{SSE}/(N-2)}{\text{SST}/(N-1)}$$

and

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

- Implying

$$r_a^2 = 1 - \frac{N-1}{N-2}(1 - r^2)$$

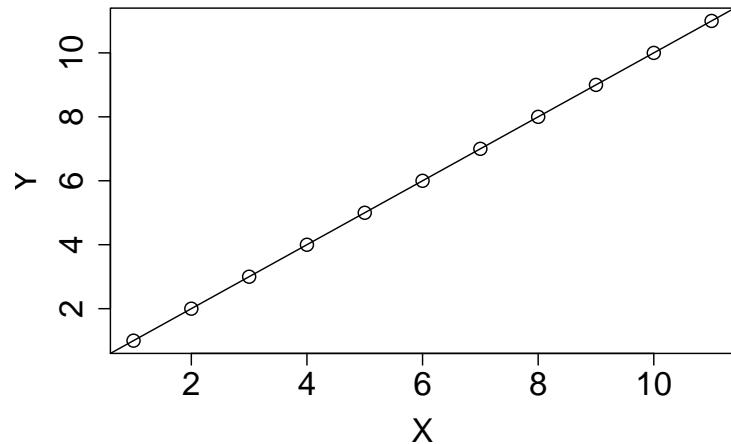
- Thus $r^2 \approx r_a^2$ for large N

Unadjusted r^2

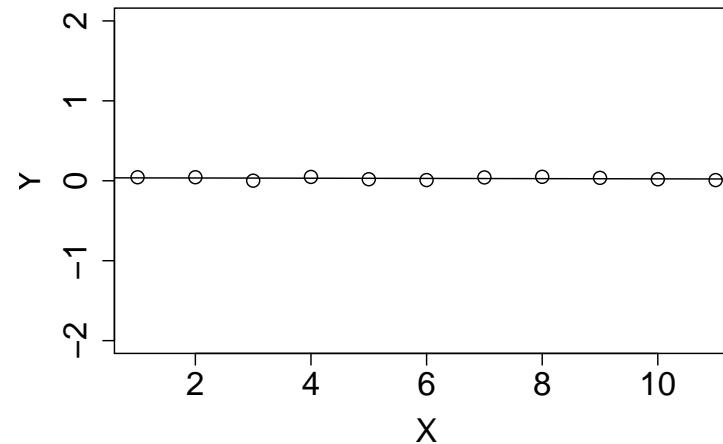
- Proportion of total variation attributable to regression
- Degree of linear association
- Ranges between 0 and 1
- $r^2 = 0 \Rightarrow$ no linear association between X and Y;
however, a non-linear association may still exist!
- $r^2 = 1$ indicates perfect fit; assessment of fit also by diagnostics
- $r^2 = 1 - \text{SSE}/\text{SST}$ typically increases with range/spacing of X

Examples of r^2

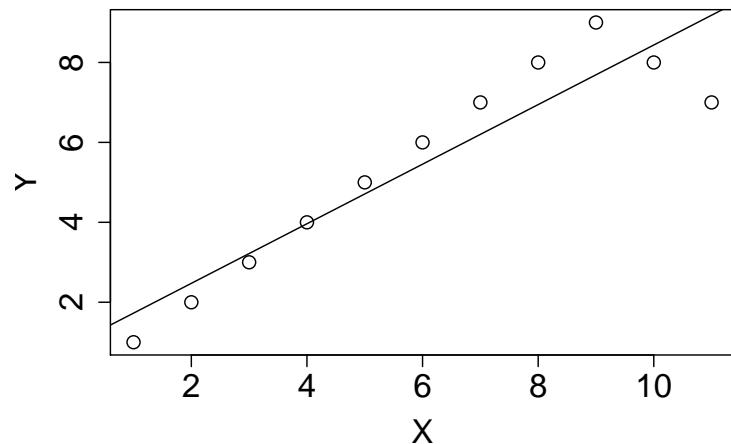
$r^2 = 1.00$



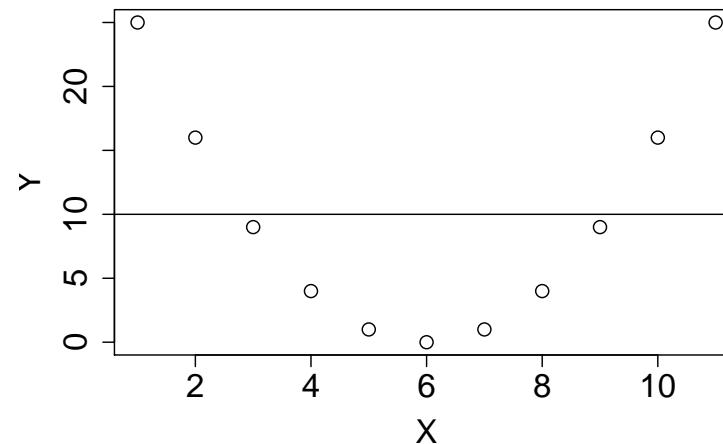
$r^2 = 0.06$



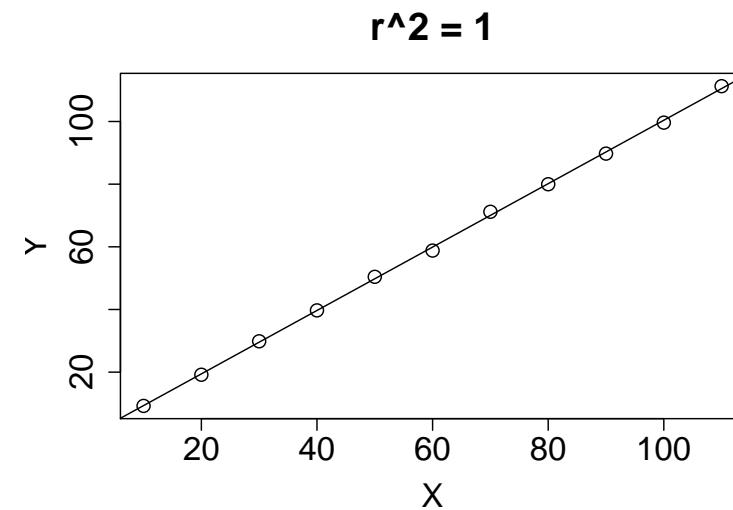
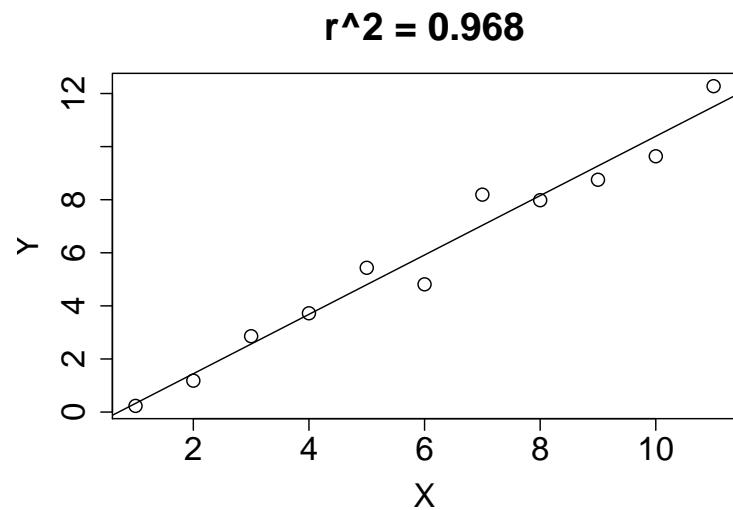
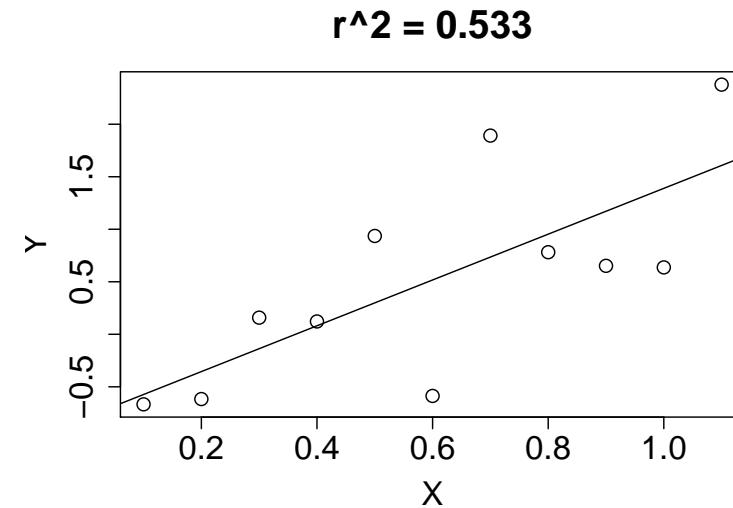
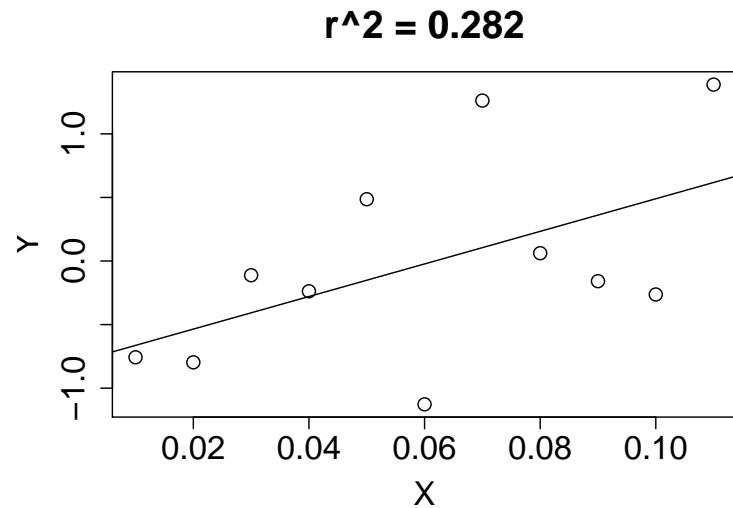
$r^2 = 0.86$



$r^2 = 0$

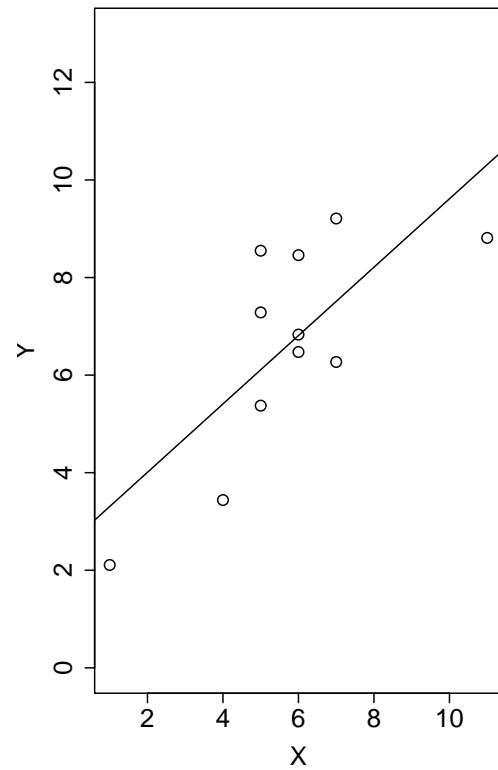


Examples of r^2 : $Y = 0 + 1 \cdot X + \epsilon$, $\epsilon \sim N(0, 1)$

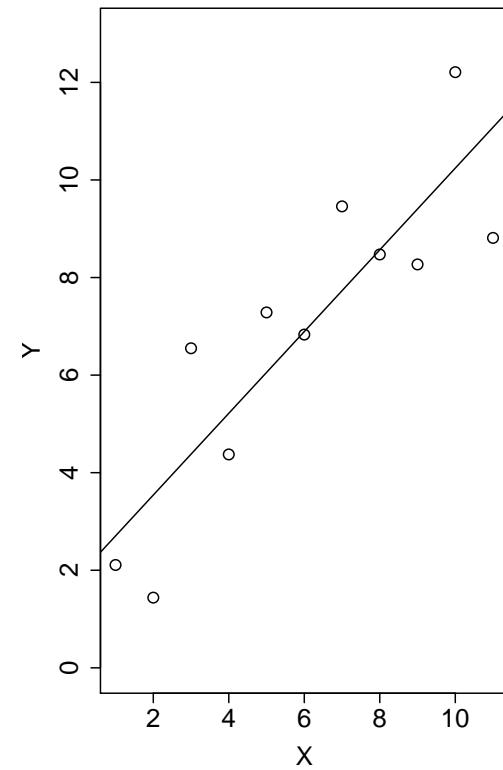


Examples of r^2 : $Y = 0 + 1 \cdot X + \epsilon$, $\epsilon \sim N(0, 4)$

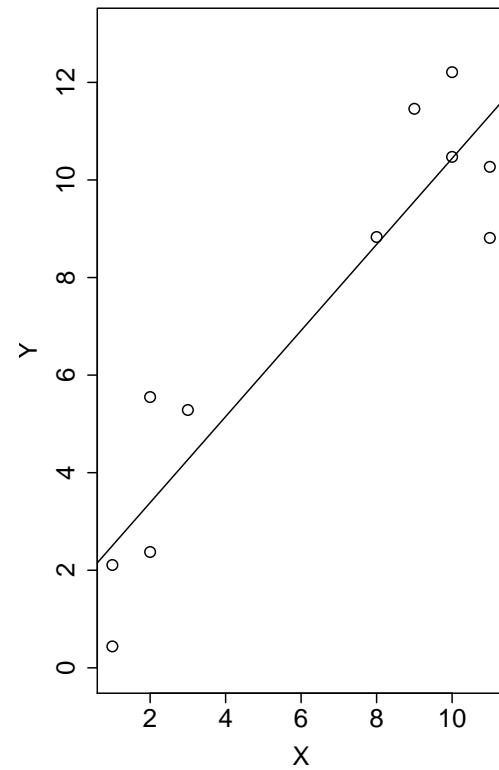
$r^2 = 0.56$, $\text{Var}(x) = 5.8$



$r^2 = 0.76$, $\text{Var}(x) = 11$



$r^2 = 0.85$, $\text{Var}(x) = 18.6$



BIOS 662 Fall 2018

Linear Regression, Part II

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- ANOVA
- Matrix formulation
- Two-sample t-test
- Diagnostics
- Measurement error

Analysis of Variance

- Recall that under $H_0 : \beta = 0$,

$$t = \frac{\hat{\beta}}{\sqrt{s_{y.x}^2 / \sum_i (X_i - \bar{X})^2}} \sim t_{N-2}$$

- Equivalently,

$$t = \frac{[XY]/[X^2]}{\sqrt{s_{y.x}^2/[X^2]}} \sim t_{N-2}$$

- In general, if $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$. Thus

$$t^2 = \frac{[XY]^2/[X^2]}{s_{y.x}^2} \sim F_{1,N-2}$$

Analysis of Variance

- Note

$$\begin{aligned}\text{SSR} &= \sum (\hat{Y}_i - \bar{Y})^2 = \sum (\hat{\alpha} + \hat{\beta}X_i - \bar{Y})^2 \\ &= \sum (\bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}X_i - \bar{Y})^2 \\ &= \sum \hat{\beta}^2(X_i - \bar{X})^2 \\ &= \frac{[XY]^2}{[X^2]^2} \sum (X_i - \bar{X})^2 = \frac{[XY]^2}{[X^2]}\end{aligned}$$

- Thus

$$t^2 = \frac{\text{SSR}}{\text{MSE}} = \frac{\text{SSR}}{\text{SSE}/(N - 2)}$$

Analysis of Variance

- If $\beta = 0$ then

$$\frac{\text{SSR}}{\sigma^2} \sim \chi_1^2 \quad \perp \quad \frac{\text{SSE}}{\sigma^2} \sim \chi_{N-2}^2$$

(*Cochran's theorem*: cf. Neter et al. p.76, 1996)

- Thus

$$t^2 = \frac{\text{SSR}/1}{\text{SSE}/(N-2)} \sim F_{1,N-2}$$

Analysis of Variance

- For $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$, we can use F with

$$C_\alpha = \{F : F > F_{1,N-2;1-\alpha}\}$$

- For the two-sided alternative the F and t tests are equivalent
- For a one-sided alternative, use t

Analysis of Variance

- ANOVA table:

Source	df	SS	MS	F
Regression	1	SSR	SSR	MSR/MSE
Residual	$N - 2$	SSE	$SSE/(N - 2)$	
Total	$N - 1$	SST		

Matrix Formulation

- Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

- Linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Matrix Formulation

- Equations (1) and (2) from previous set of notes:

$$-\bar{Y} + \alpha + \beta \bar{X} = 0$$

$$-\sum_i X_i Y_i + \alpha \sum_i X_i + \beta \sum_i X_i^2 = 0$$

- Equivalent to:

$$\mathbf{X}' \mathbf{X} \boldsymbol{\beta} = \mathbf{X}' \mathbf{Y}$$

Matrix Formulation

- Therefore

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- We can also show

$$\text{SST} = \mathbf{Y}'\mathbf{Y} - \frac{1}{N}\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

$$\text{SSR} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{1}{N}\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

$$\text{SSE} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$$

where \mathbf{J} is an $n \times n$ matrix of 1s

Linear Regression and Two Sample t-test

- Define

$$X = \begin{cases} 1 & \text{if in group 1} \\ 0 & \text{if in group 2} \end{cases}$$

- X is called an *indicator* or *dummy* variable
- Model

$$Y = \alpha + \beta X + \epsilon$$

Linear Regression and Two Sample t-test

- Suppose we have two groups of observations: Y_{1i} for $i = 1, \dots, n_1$ and Y_{2i} for $i = 1, \dots, n_2$
- Recall that the test statistic for the two sample t-test is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where

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

Linear Regression and Two Sample t-test

- Let

$$N = n_1 + n_2$$

$$(Y_1, \dots, Y_{n_1}) = (Y_{11}, \dots, Y_{1n_1})$$

$$(Y_{n_1+1}, \dots, Y_N) = (Y_{21}, \dots, Y_{2n_2})$$

$$X_i = \begin{cases} 1 & \text{if in group 1} \\ 0 & \text{if in group 2} \end{cases}$$

Linear Regression and Two Sample t-test

- Consider the regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i; \quad i = 1, 2, 3, \dots, N$$

- Note that

$$\begin{aligned}[X^2] &= \sum_i (X_i - \bar{X})^2 = \sum X_i^2 - N\bar{X}^2 \\ &= n_1 - N \left(\frac{n_1}{N} \right)^2 \\ &= n_1 \left(1 - \frac{n_1}{N} \right) \\ &= \frac{n_1 n_2}{N}\end{aligned}$$

Linear Regression and Two Sample t-test

- Recall that

$$\hat{\beta} = \sum c_i Y_i$$

where $c_i = (X_i - \bar{X})/[X^2]$

- Thus

$$\begin{aligned}\hat{\beta} &= \frac{(1 - \bar{X}) \sum_{i=1}^{n_1} Y_i}{[X^2]} + \frac{(-\bar{X}) \sum_{i=n_1+1}^N Y_i}{[X^2]} \\ &= \bar{Y}_1 - \bar{Y}_2\end{aligned}$$

- We can show that

$$s_{y \cdot x}^2 = s_p^2$$

Linear Regression and Two Sample t-test

- Therefore:

$$\begin{aligned} t &= \frac{\hat{\beta}}{\sqrt{s_{y \cdot x}^2 / \sum_i (X_i - \bar{X})^2}} \\ &= \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{N / (n_1 n_2)}} \end{aligned}$$

Linear Regression and Two Sample t-test

- Example: Body fat in Native American children
- Percent body fact (PBF) measured by bioelectric impedance and skinfold thickness
- Two tribes: Apache (mountains) and Tohona (desert)
- Question: Is the mean PBF the same in Apache and Tohona children?
- Samples: Tohona ($n = 63$); Apache ($n = 35$)

Linear Regression and Two Sample t-test

- Two sample t-test:

```
proc ttest;  
  var pbf;  
  class tribe;
```

The TTEST Procedure

Variable: pbf

tribe	N	Mean	Std Dev	Std Err
Apache	35	33.1757	6.9215	1.1700
Tohona	63	37.3615	8.0349	1.0123
Diff (1-2)		-4.1857	7.6591	1.6147

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	96	-2.59	0.0110
Satterthwaite	Unequal	79.523	-2.71	0.0083

Linear Regression and Two Sample t-test

- Model

$$Y = \alpha + \beta X + \epsilon$$

where

$$Y = \text{PBF}$$

and

$$X = \begin{cases} 1 & \text{if Apache} \\ 0 & \text{if Tohona} \end{cases}$$

Linear Regression and Two Sample t-test

```
proc reg;  
    model pbf=apache;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	394.20974	394.20974	6.72	0.0110
Error	96	5631.59441	58.66244		
Corrected Total	97	6025.80415			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	37.36147	0.96496	38.72	<.0001
apache	1	-4.18574	1.61469	-2.59	0.0110

Diagnostics

- Assumptions for linear regression

1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$

2. X s are fixed constants

3. ϵ_i iid $\sim N(0, \sigma^2)$

(homogeneity of variance)

- *Residual plot:* Scatterplot of

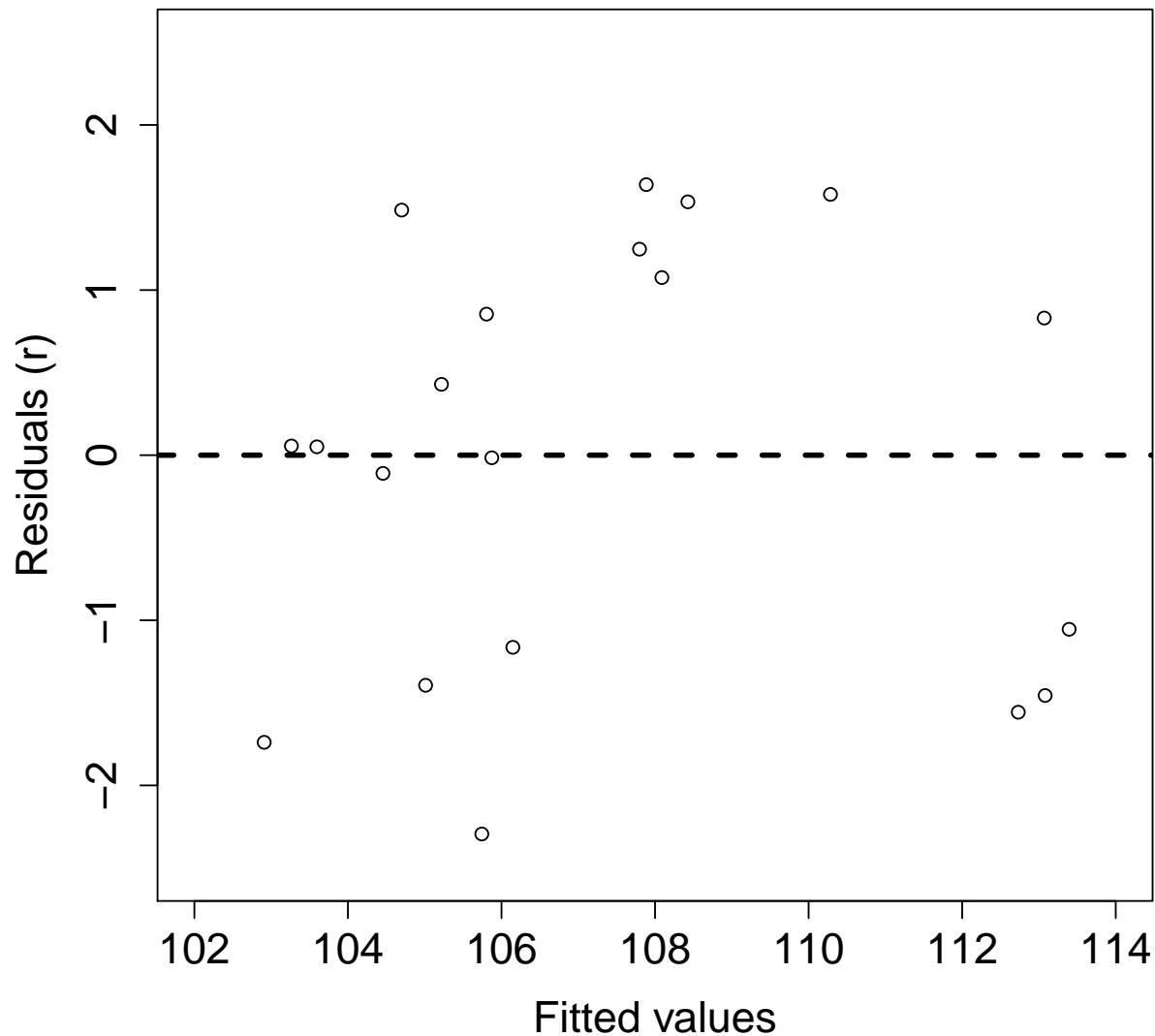
$$(\hat{Y}_i, r_i) = (\hat{Y}_i, Y_i - \hat{Y}_i)$$

- If we see lack of homogeneity of variance or of linearity, consider transformations; see Table 10.28 (page 399) of the text

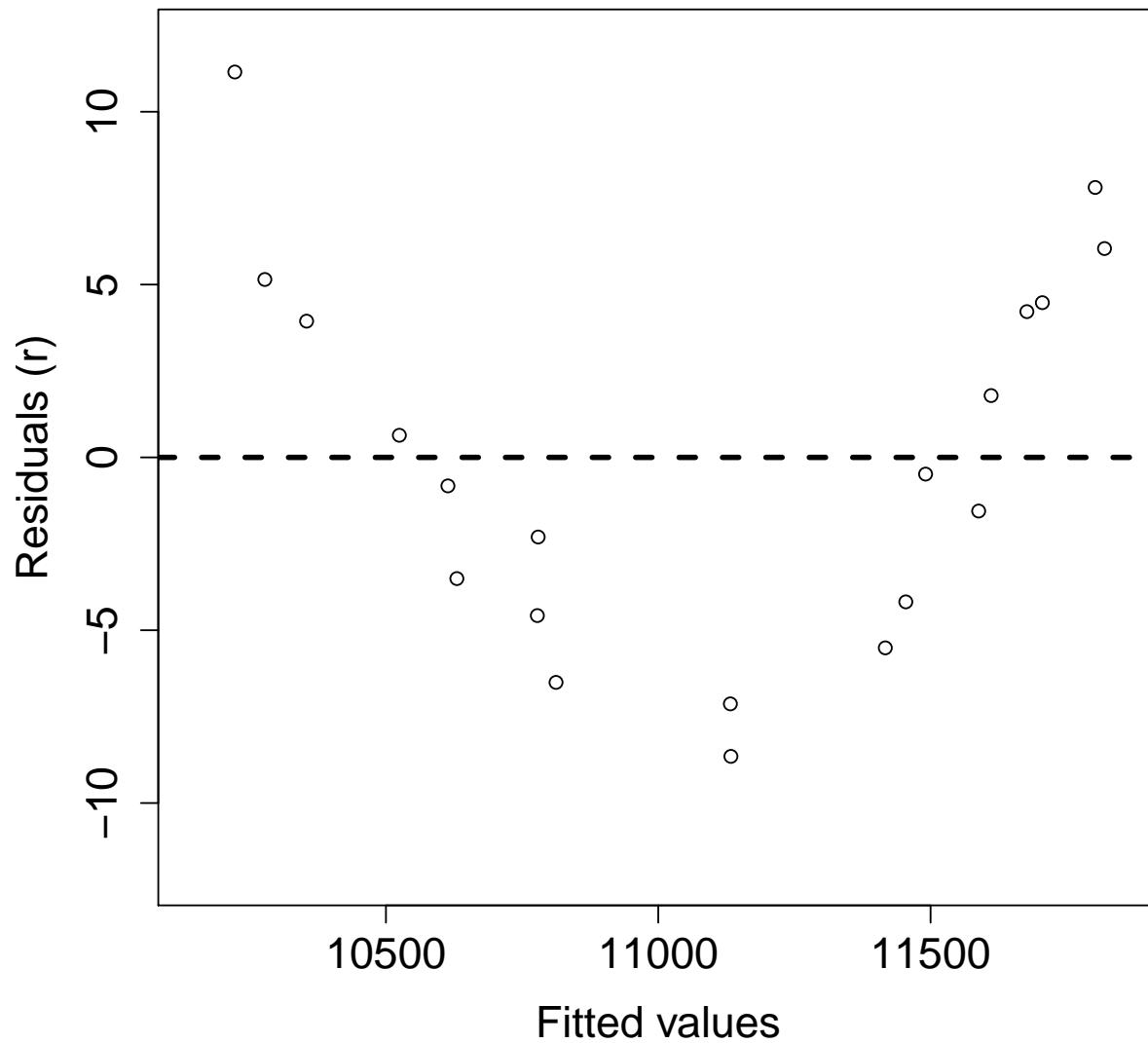
Diagnostics

- The following three pages contain prototypical residual plots indicating successively:
 1. linear regression model is appropriate
 2. assumption of linearity questionable
 3. assumption of constant variance questionable

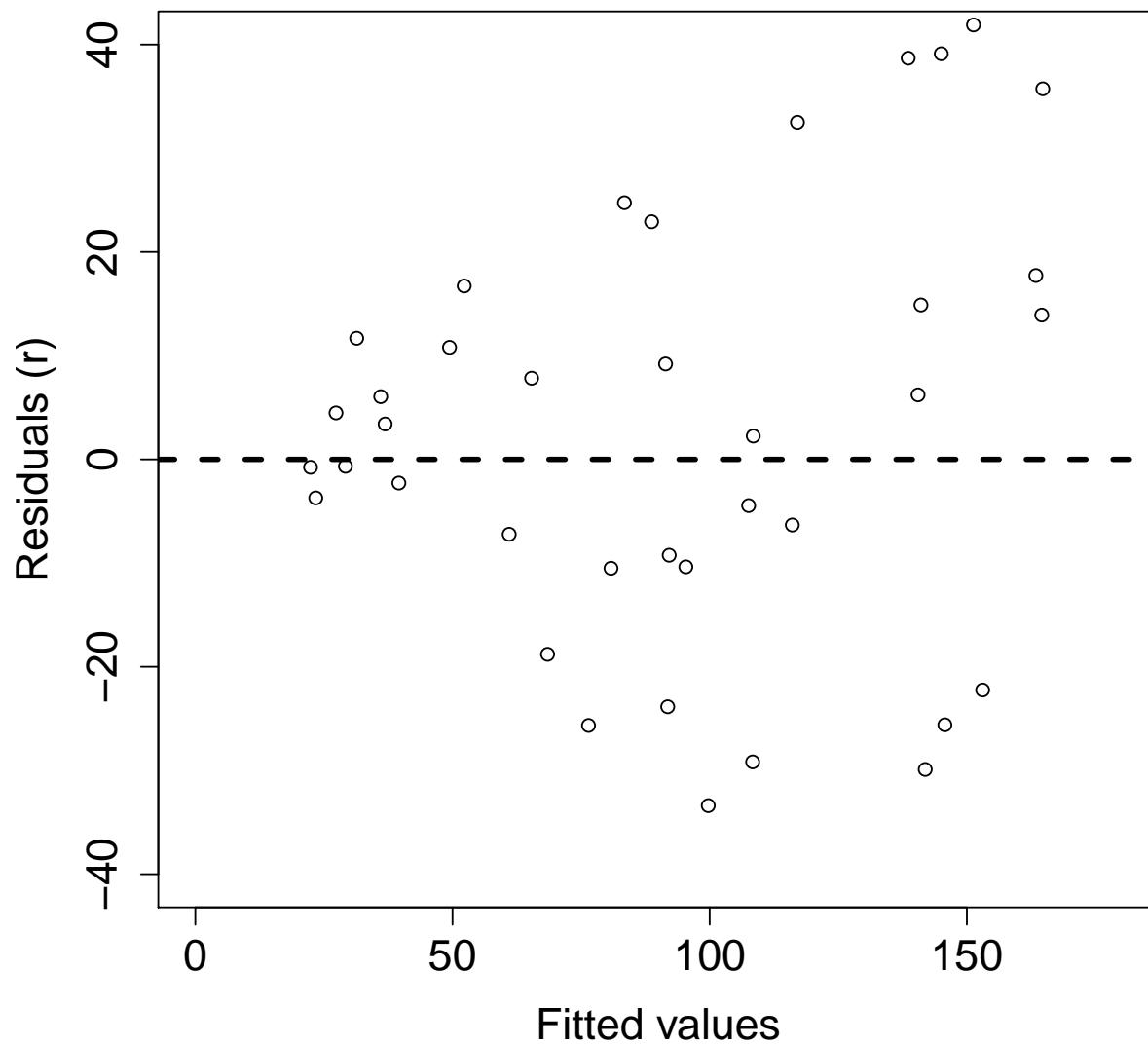
Regression: Residuals



Regression: Residuals



Regression: Residuals



Regression: Example

- FEV_1 as a function of age in male children

```
proc reg;  
  model fev1=age;
```

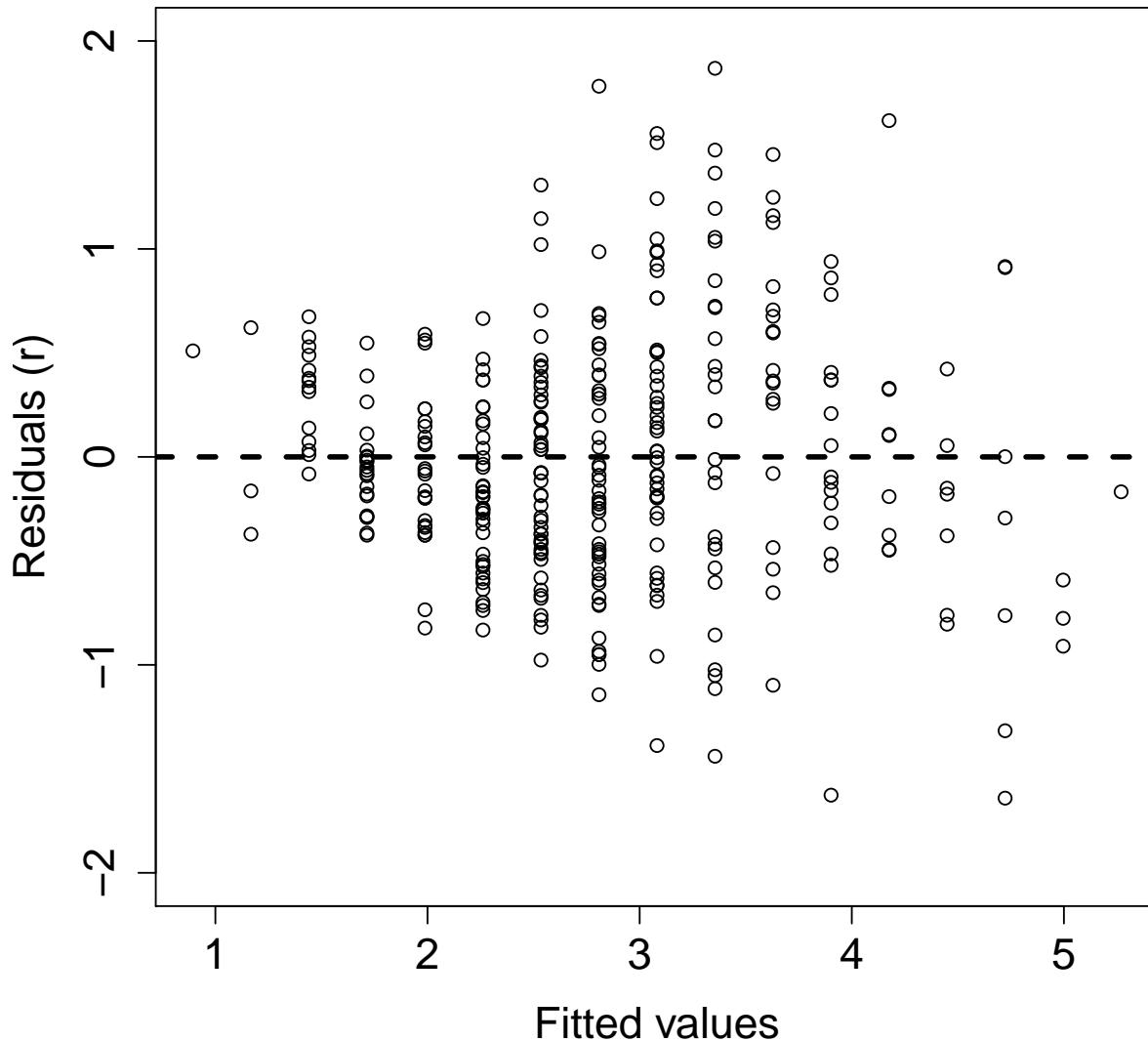
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	221.89640	221.89640	641.57	<.0001
Error	334	115.51840	0.34586		
Corrected Total	335	337.41480			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.07360	0.11279	0.65	0.5145
age	1	0.27348	0.01080	25.33	<.0001

Regression: Example cont.



Regression: Example cont.

- Regress $\log(\text{FEV}_1)$ on age for male children

```
proc reg;  
  model logfev1=age;
```

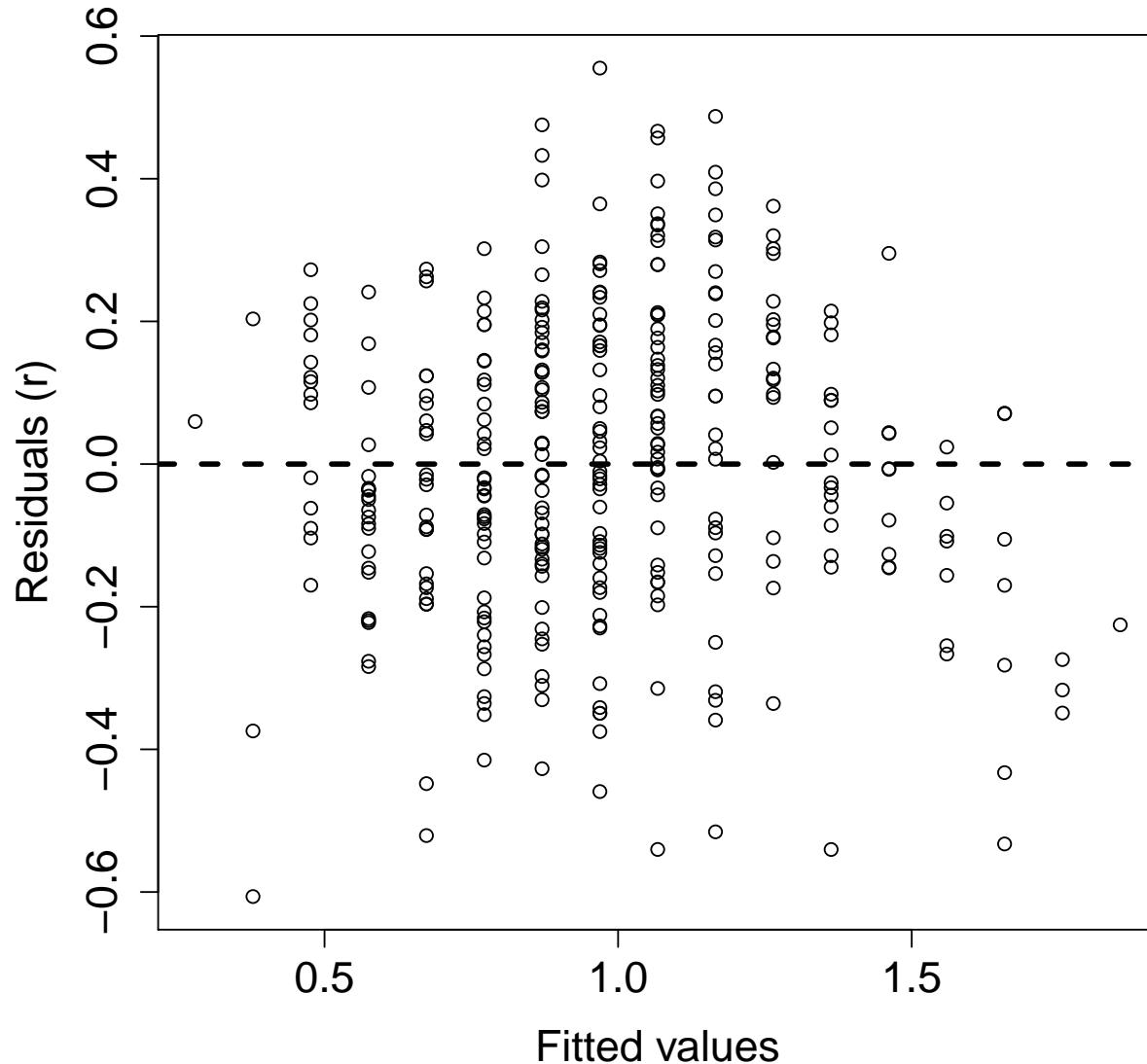
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	28.76362	28.76362	651.53	<.0001
Error	334	14.74543	0.04415		
Corrected Total	335	43.50906			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.01569	0.04030	-0.39	0.6973
age	1	0.09846	0.00386	25.53	<.0001

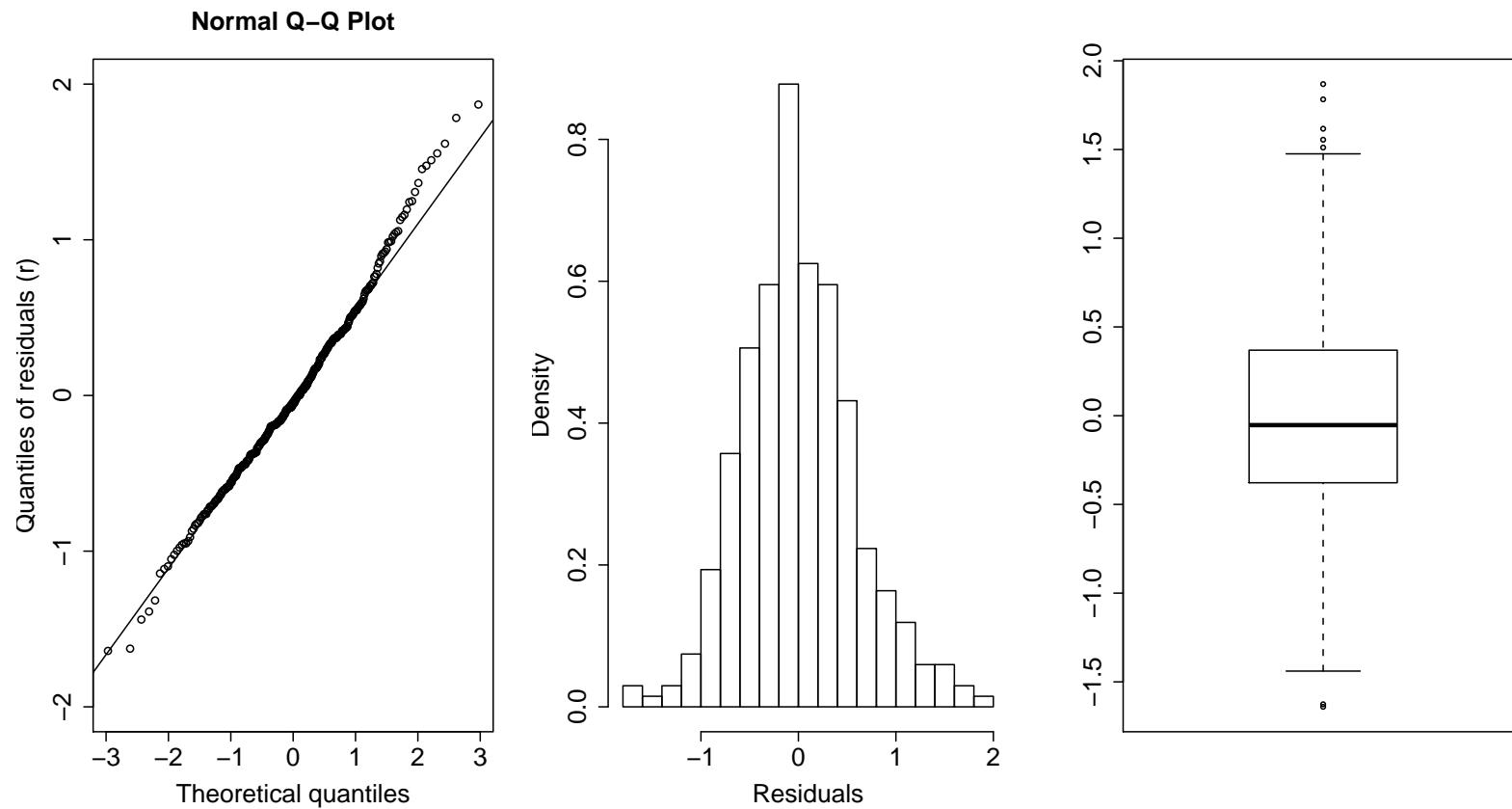
Regression: Example cont.



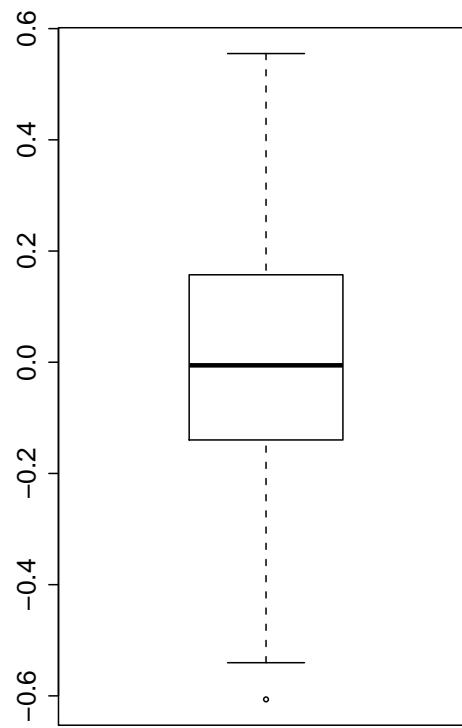
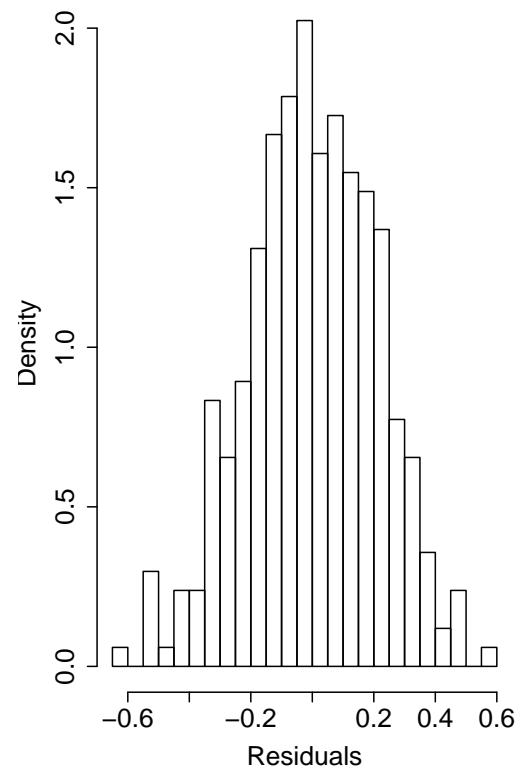
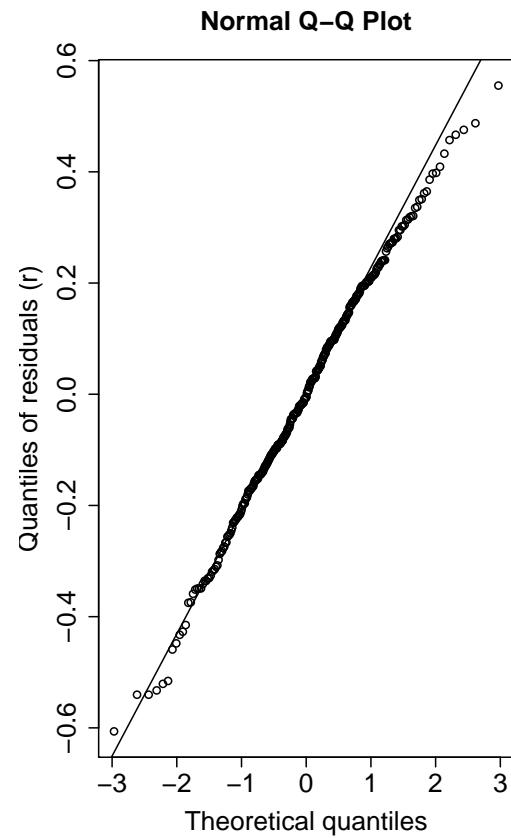
Normality Diagnostics

- Assumption: The ϵ_i are normally distributed
- This assumption is not as important if N is large (CLT)
- Inference robust to small departures from normality
- Violations of other assumptions can suggest non-normality
- Tests of normality of residuals; beware lack of power
- qq-plot, histogram, boxplot of residuals

Normality Diagnostics: FEV₁

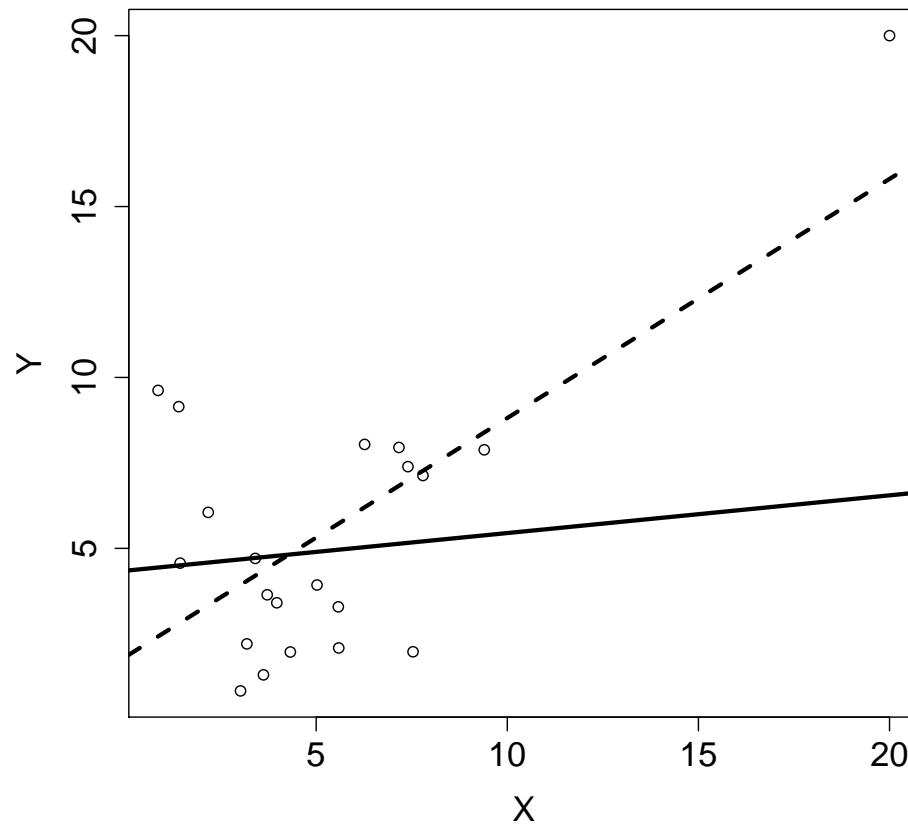


Normality Diagnostics: $\log(\text{FEV}_1)$



Regression: Diagnostics

- Beware influential observations; always check scatterplot



Regression: Graphical Diagnostics in SAS

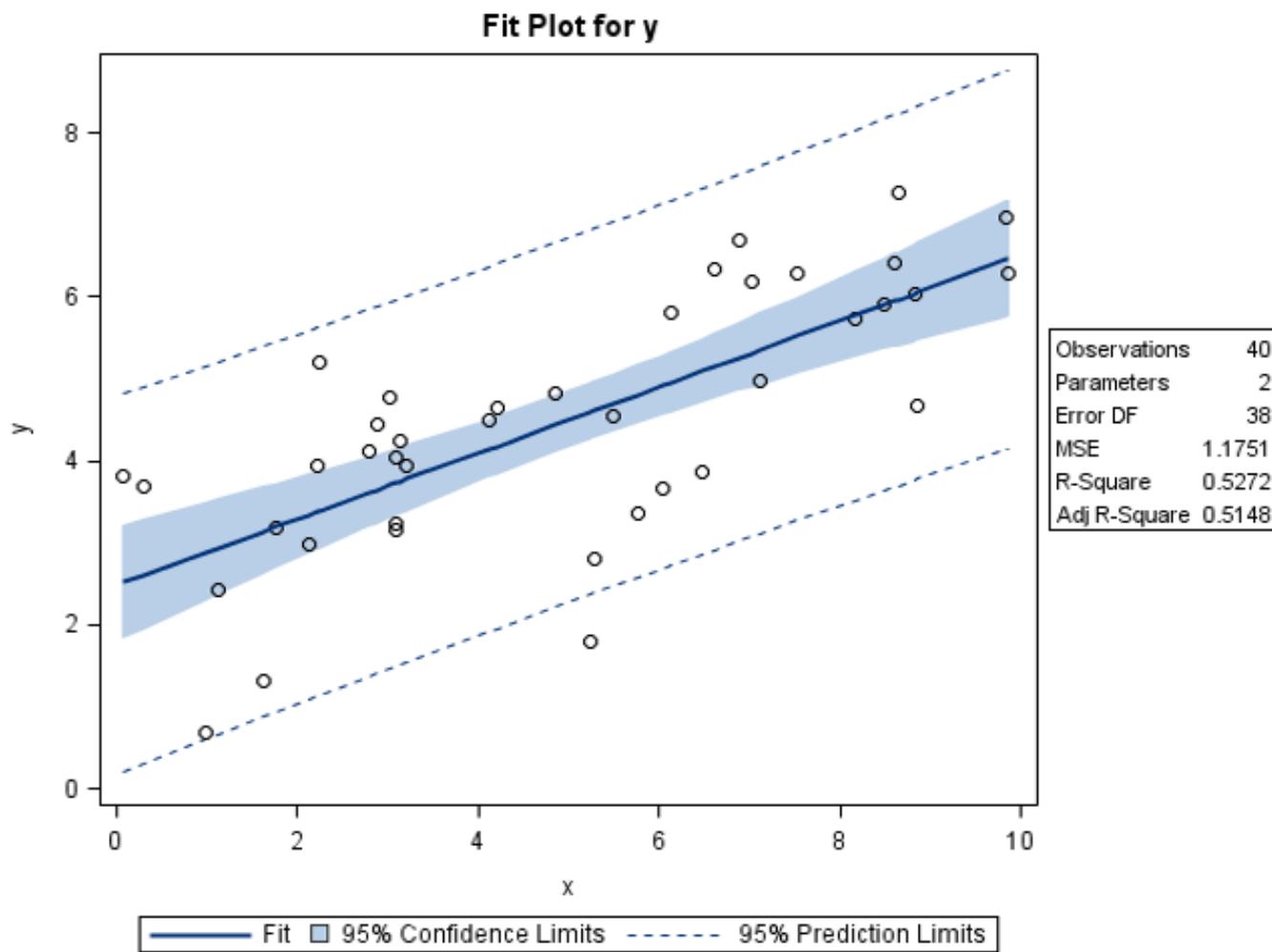
- Use ODS graphics in SAS 9.2 or later
- Default plots often sufficient, use options in plots= to specify particular plots

```
ods graphics on;
ods rtf file='diagnostics.rtf';

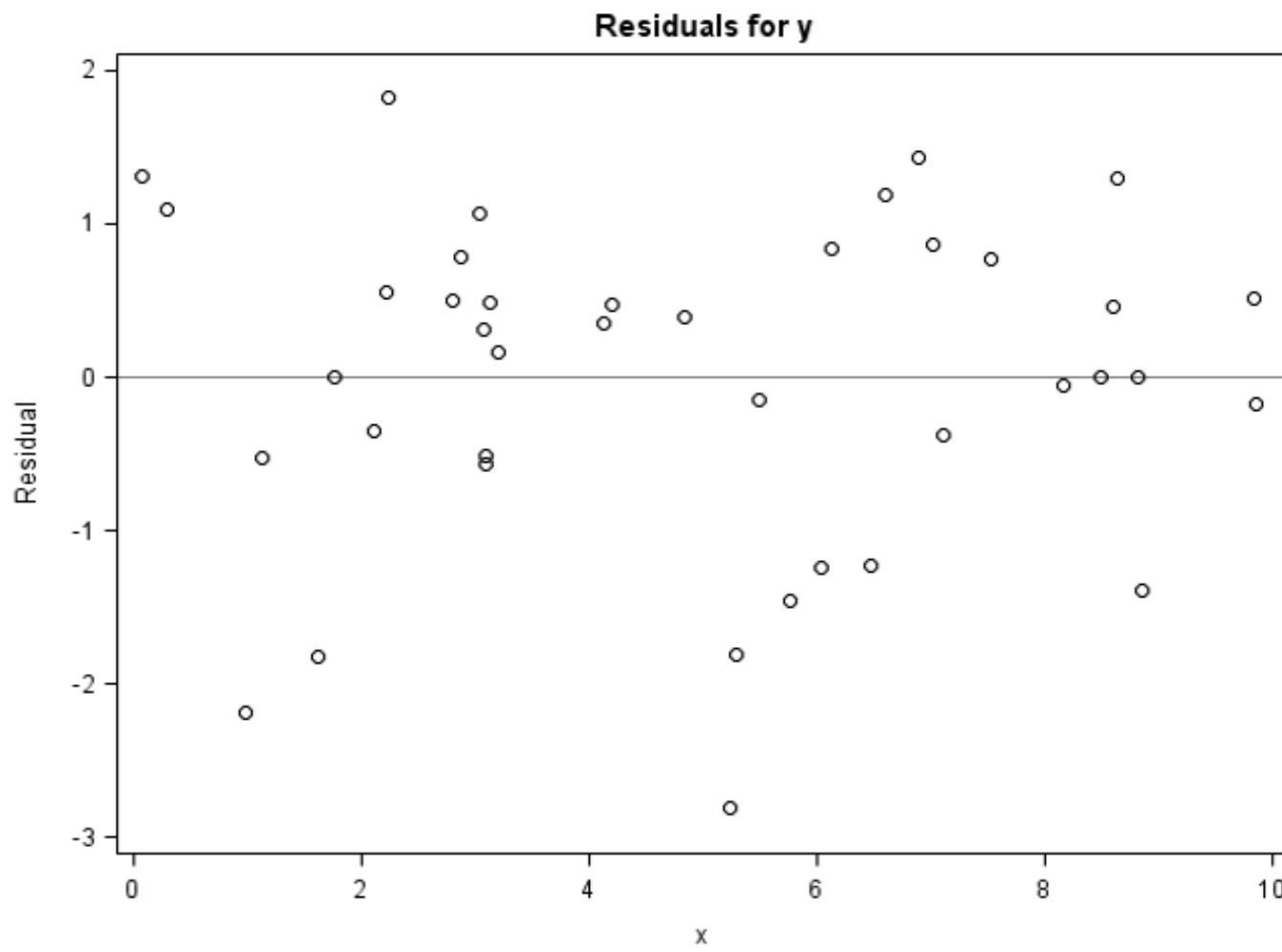
proc reg data=diagnostics;
  model y = x;

run; quit;
ods rtf close;
```

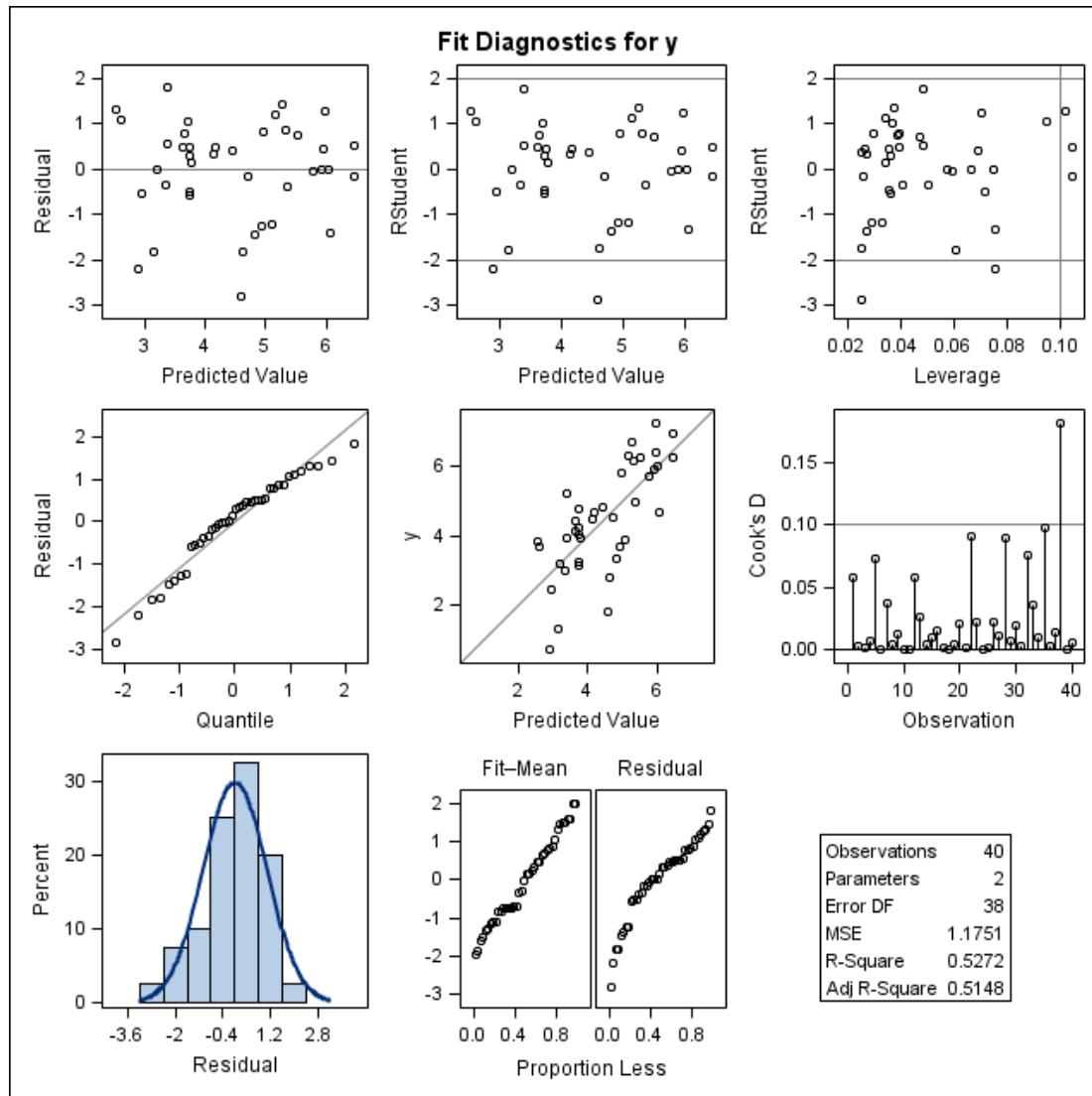
Regression: Graphical Diagnostics in SAS



Regression: Graphical Diagnostics in SAS



Regression: Graphical Diagnostics in SAS



Remedial Measures

- Transformations, e.g., $\log(Y) = \alpha + \beta X$
- Multiple regression, e.g., $Y = \alpha + \beta_1 X + \beta_2 X^2$
- Nonparametric procedures, e.g., Kendall's tau
- More sophisticated models allowing for
 - dependencies/clusters (e.g., GEE)
 - heterogeneity of variance (e.g., weighted least squares)

Regression: X Random

- Assumption: X s are known
- Suppose X and Y are both random variables

$$Y = \alpha + \beta_{y \cdot x} X + \epsilon$$

$$X \perp \epsilon; \text{ Var}(X) = \delta^2$$

- Results on estimation, testing, and prediction still hold (Neter et al., 1996 p 85; Section 2.9.2 of Abraham and Ledolter, 2006)
- The covariance between two random variables X and Y is defined as

$$\text{Cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$$

Regression: X Random

- Now

$$\beta_{y \cdot x} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

- Proof: We have $\text{Cov}(a + bW, U) = b\text{Cov}(W, U)$
and $\text{Cov}(W, U + V) = \text{Cov}(W, U) + \text{Cov}(W, V)$
- Thus

$$\begin{aligned}\text{Cov}(Y, X) &= \text{Cov}(\alpha + \beta_{y \cdot x}X + \epsilon, X) \\ &= \beta_{y \cdot x}\text{Cov}(X, X) + \text{Cov}(\epsilon, X) \\ &= \beta_{y \cdot x}\text{Var}(X)\end{aligned}$$

Measurement Error

- Instead of observing X , we observe

$$W = X + U$$

where U is a random variable with

$$E(U) = 0, \quad \text{Var}(U) = \tau^2$$

$$U \perp X, \quad U \perp Y$$

- Then

$$\text{Cov}(W, Y) = \text{Cov}(X + U, Y)$$

$$= \text{Cov}(X, Y) + \text{Cov}(U, Y) = \text{Cov}(X, Y)$$

Measurement Error

- By independence

$$\text{Var}(W) = \text{Var}(X) + \text{Var}(U) = \delta^2 + \tau^2$$

- Thus

$$\begin{aligned}\beta_{y \cdot w} &= \frac{\text{Cov}(Y, W)}{\text{Var}(W)} \\ &= \frac{\text{Cov}(Y, X)}{\delta^2 + \tau^2} \\ &= \frac{\delta^2}{\delta^2 + \tau^2} \frac{\text{Cov}(Y, X)}{\delta^2} \\ &= \frac{\delta^2}{\delta^2 + \tau^2} \beta_{y \cdot x}\end{aligned}$$

Measurement Error

- Because

$$0 \leq \frac{\delta^2}{\delta^2 + \tau^2} \leq 1,$$

it follows that

$$|\beta_{y \cdot w}| \leq |\beta_{y \cdot x}|$$

- That is, there is attenuation towards the null

Measurement Error

- Thus if X is not determined precisely, we underestimate the strength of association between X and Y
- Reliability coefficient of X :

$$R_{\text{rel}} = \frac{\delta^2}{\delta^2 + \tau^2}$$

- If R_{rel} is known,

$$\tilde{\beta} = R_{\text{rel}}^{-1} \hat{\beta}_{y \cdot w}$$

is an unbiased estimator of $\beta_{y \cdot x}$

Measurement Error

- Because

$$\text{Var}(\tilde{\beta}) = R_{\text{rel}}^{-2} \text{Var}(\hat{\beta}_{y \cdot w})$$

the t -statistic for testing $H_0 : \beta_{y \cdot x} = 0$ is

$$t_{y \cdot x} = \frac{\tilde{\beta}}{\sqrt{\text{Var}(\tilde{\beta})}} = \frac{R_{\text{rel}}^{-1} \hat{\beta}_{y \cdot w}}{\sqrt{R_{\text{rel}}^{-2} \text{Var}(\hat{\beta}_{y \cdot w})}} = t_{y \cdot w}$$

Measurement Error

- Suppose there are k independent measures of W made on each person in a study
- It can be shown that

$$\text{Var}(\bar{W}_k) = \delta^2 + \frac{\tau^2}{k}$$

- Therefore

$$\beta_{y \cdot \bar{w}_k} = \frac{\delta^2}{\delta^2 + \tau^2/k} \beta_{y \cdot x} \rightarrow \beta_{y \cdot x} \quad \text{as } k \rightarrow \infty$$

Measurement Error

- For example, suppose W is a physiological variable such as BP or cholesterol
- If we get two or more measures of W , the bias will be reduced
- For cholesterol, $R_{\text{rel}} \approx 0.8$ and $\delta^2 + \tau^2 \approx 1600$
- Therefore

$$\tau^2 = 0.2(1600) = 320$$

- If $k = 2$, $1280/(1280 + 320/2) = 0.89$
If $k = 3$, $1280/(1280 + 320/3) = 0.92$

Measurement Error

- Measurement error is likely to be present in most situations; however, it is usually ignored because:
 - Often practically negligible (e.g., if can use precise instrumentation)
 - Interest is in inference/prediction based on observable random variables
- Random measurement error in Y is absorbed into ϵ

BIOS 662 Fall 2018

Linear Regression, Part III

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Multiple linear regression
- Measures of association
- Parametric/large N
 - Pearson correlation coefficient
- Nonparametric (i.e., rank based)
 - Spearman rank correlation coefficient
 - Kendall's τ

Multiple Linear Regression

Reasons for using multiple linear regression rather than just simple linear regression include:

- Determining the best set of variables with which to predict an outcome variable
- Allowing adjustment for potential confounders when investigating an exposure–disease association
- Investigating potential interactions between exposures associated with a disease
- Using a categorical predictor with more than two categories

Some of these reasons may apply simultaneously

Multiple Linear Regression Model

- Multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i,$$

$$i = 1, 2, \dots, N$$

- Data are (Y_i, \mathbf{X}_i) ; $i = 1, 2, \dots, N$, where \mathbf{X}_i is a vector of length k
- Assumptions:

1. Linearity: each X variable is linearly associated with Y
2. The values of each X variable are fixed constants
3. ϵ_i iid $N(0, \sigma^2)$

Multiple Linear Regression Model

Multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i,$$
$$i = 1, 2, \dots, N$$

Interpretation of parameters:

- β_j is the change in the expected value of Y when the j^{th} X variable increases by one unit, with all the other X variables being held constant
- If the j^{th} X variable is dichotomous, that is, takes on only values in $\{0, 1\}$, this corresponds to the difference between $E(Y)$ when the value of the j^{th} X is 1 versus when it is 0

Matrix Formulation

- Let

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Matrix Formulation

- Linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- The least squares estimators are the solutions to the set of equations:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$$

- Therefore, as in the simple linear regression case:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- The coefficient of determination is now written as R^2 (rather than r^2); as before it is the proportion of the total variation attributable to regression (that is, explained by all the X variables together)

Analysis of Variance

- ANOVA table:

Source	df	SS	MS	F
Regression	k	SSR	MSR = SSR/k	MSR/MSE
Residual	$N - k - 1$	SSE	MSE = SSE/(N - k - 1)	
Total	$N - 1$	SST		

- The F test is for

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

versus

$$H_A : \text{at least one } \beta_j \neq 0$$

Multiple Linear Regression Example

- Consider the SBP and age example and suppose we want to investigate whether the association varies with gender
- Let:
 - Y_i be the systolic blood pressure of person i
 - X_{i1} be the age of person i
 - X_{i2} be 1 if person i is male and 0 otherwise
 - $X_{i3} = X_{i1} \cdot X_{i2}$

Multiple Linear Regression Example

```
proc reg;  
    model sbp = age male;
```

Dependent Variable: sbp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2414.90795	1207.45397	288.31	<.0001
Error	37	154.95563	4.18799		
Corrected Total	39	2569.86358			

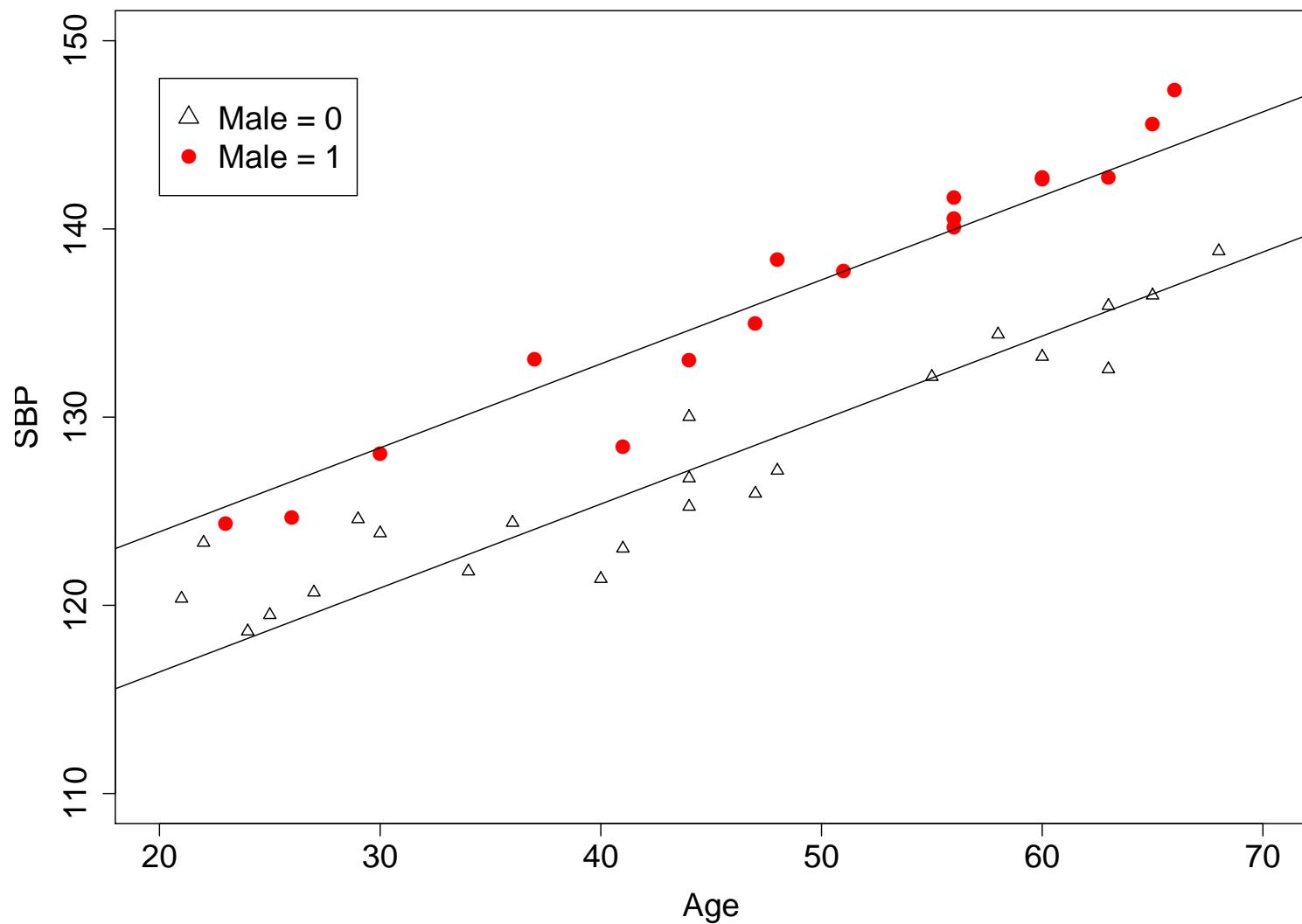
Multiple Linear Regression Example

Root MSE	2.04646	R-Square	0.9397
Dependent Mean	131.15651	Adj R-Sq	0.9364
Coeff Var	1.56032		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	107.52982	1.08737	98.89	<.0001
age	1	0.44634	0.02268	19.68	<.0001
male	1	7.44864	0.65488	11.37	<.0001

Multiple Linear Regression Example



Multiple Linear Regression Example

```
proc reg;  
    model sbp = age male agemale;
```

Dependent Variable: sbp

Number of Observations Read	40
Number of Observations Used	40

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2445.33277	815.11092	235.64	<.0001
Error	36	124.53081	3.45919		
Corrected Total	39	2569.86358			

Multiple Linear Regression Example

Root MSE	1.85989	R-Square	0.9515
Dependent Mean	131.15651	Adj R-Sq	0.9475
Coeff Var	1.41807		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	109.92861	1.27705	86.08	<.0001
age	1	0.39170	0.02765	14.17	<.0001
male	1	1.81501	1.99065	0.91	0.3680
agemale	1	0.12305	0.04149	2.97	0.0053

Multiple Linear Regression Example

```
> fit <- lm(sbp~age+male+agemale)
> summary(fit)
```

Call:

```
lm(formula = sbp ~ age + male + agemale)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	109.92860	1.27708	86.078	< 2e-16 ***
age	0.39170	0.02765	14.168	2.7e-16 ***
male	1.81503	1.99070	0.912	0.36797
agemale	0.12305	0.04149	2.966	0.00533 **

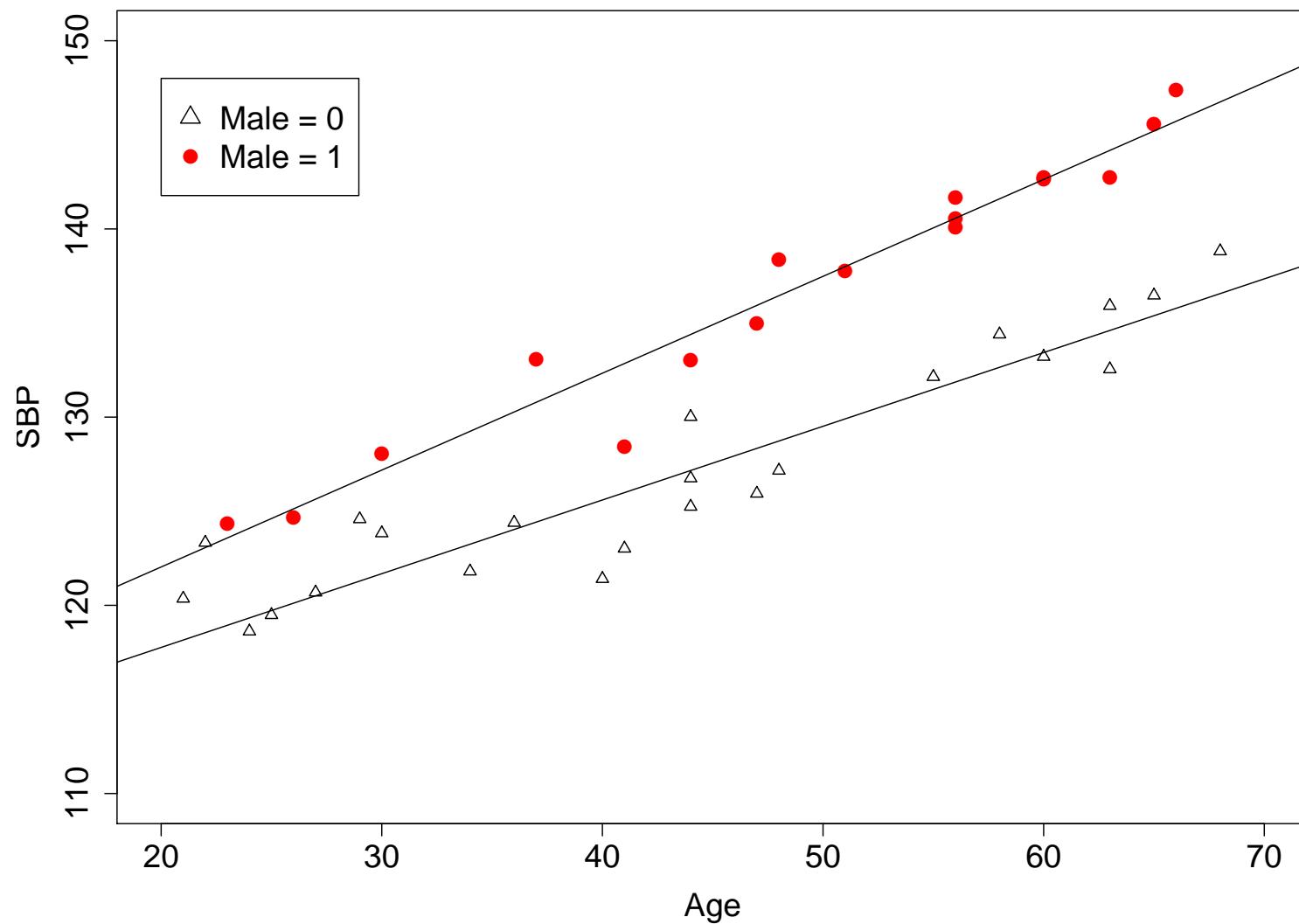
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.86 on 36 degrees of freedom

Multiple R-squared: 0.9515, Adjusted R-squared: 0.9475

F-statistic: 235.6 on 3 and 36 DF, p-value: < 2.2e-16

Multiple Linear Regression Example



Multiple Linear Regression Example

Now suppose we use age in 10-year age groups

```
data sbp;
  set sbp;

agegroup=10*floor(age/10);

if 20 le age lt 30 then age2029=1;
  else age2029=0;
if 30 le age lt 40 then age3039=1;
  else age3039=0;
if 40 le age lt 50 then age4049=1;
  else age4049=0;
if 50 le age lt 60 then age5059=1;
  else age5059=0;
if 60 le age lt 70 then age6069=1;
  else age6069=0;
```

Multiple Linear Regression Example

```
proc reg;  
model sbp = agegroup;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1785.32369	1785.32369	86.47	<.0001
Error	38	784.53989	20.64579		
Corrected Total	39	2569.86358			

Root MSE	4.54376	R-Square	0.6947
Dependent Mean	131.15651	Adj R-Sq	0.6867
Coeff Var	3.46438		

Multiple Linear Regression Example

Parameter Estimates

Variable	DF	Parameter	Standard	t Value	Pr > t
		Estimate	Error		
Intercept	1	111.95282	2.18650	51.20	<.0001
agegroup	1	0.46554	0.05006	9.30	<.0001

Assumption here: SBP changes by the same amount from each age group to the next.

Multiple Linear Regression Example

```
proc reg;  
    model sbp = age3039 age4049 age5059 age6069;
```

Analysis of Variance

Source	DF	Sum of	Mean	F Value	Pr > F
		Squares	Square		
Model	4	1873.06457	468.26614	23.52	<.0001
Error	35	696.79901	19.90854		
Corrected Total	39	2569.86358			

Root MSE	4.46190	R-Square	0.7289
Dependent Mean	131.15651	Adj R-Sq	0.6979
Coeff Var	3.40197		

Multiple Linear Regression Example

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	122.01318	1.57752	77.34	<.0001
age3039	1	4.21824	2.54367	1.66	0.1062
age4049	1	6.56457	2.07327	3.17	0.0032
age5059	1	15.76066	2.40970	6.54	<.0001
age6069	1	17.78679	2.11646	8.40	<.0001

Multiple Linear Regression Example

```
proc reg;  
model sbp = age2029 age3039 age4049 age5059;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	139.79997	1.41098	99.08	<.0001
age2029	1	-17.78679	2.11646	-8.40	<.0001
age3039	1	-13.56855	2.44388	-5.55	<.0001
age4049	1	-11.22222	1.94954	-5.76	<.0001
age5059	1	-2.02613	2.30411	-0.88	0.3852

Multiple Linear Regression Example

```
proc glm;  
  class agegroup;  
  model sbp = agegroup / solution;  
  lsmeans agegroup;
```

The GLM Procedure

Class Level Information

Class	Levels	Values
-------	--------	--------

agegroup	5	20 30 40 50 60
----------	---	----------------

Number of Observations Read	40
-----------------------------	----

Number of Observations Used	40
-----------------------------	----

Multiple Linear Regression Example

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		139.7999661 B	1.41097637	99.08	<.0001
agegroup	20	-17.7867909 B	2.11646455	-8.40	<.0001
agegroup	30	-13.5685533 B	2.44388276	-5.55	<.0001
agegroup	40	-11.2222160 B	1.94954402	-5.76	<.0001
agegroup	50	-2.0261338 B	2.30411476	-0.88	0.3852
agegroup	60	0.0000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations.
Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Multiple Linear Regression Example

The GLM Procedure

Least Squares Means

agegroup	sbp LSMEAN
20	122.013175
30	126.231413
40	128.577750
50	137.773832
60	139.799966

Correlation

- The *correlation* between random variables X and Y is

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

- Note:

$$\rho = \frac{\beta_{y \cdot x} \sigma_X}{\sigma_Y} = \frac{\beta_{x \cdot y} \sigma_Y}{\sigma_X}$$

Correlation

- Estimate ρ by

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (Y_i - \bar{Y})^2 \sum_i (X_i - \bar{X})^2}} = \frac{[XY]}{\sqrt{[X^2][Y^2]}}$$

the *sample Pearson product moment correlation coefficient*

- One can show that

$$r = \hat{\beta}_{y \cdot x} \frac{s_X}{s_Y} = \text{sign}(\hat{\beta}_{y \cdot x}) \sqrt{r^2}$$

where r^2 is as in the first set of notes on regression, i.e., the proportion of total variation attributable to regression

Correlation

- The correlation coefficient r has the following properties:
 - $r \in [-1, 1]$
 - $r = 1$ iff all observations lie on a straight line with positive slope
 - $r = -1$ iff all observations lie on a straight line with negative slope
 - it is invariant under multiplication and addition of constants to X or Y
 - it measures *linear association* between two variables
 - it tends to be close to zero if there is no linear association, even if there is a strong non-linear association

Demonstrating Correlation Properties Using R

```
> x <- 1:11
```

```
> y <- x
```

```
> cor(y,x)
```

```
[1] 1
```

```
> cor(y,3*x)
```

```
[1] 1
```

```
> cor(y/100,3*x+10)
```

```
[1] 1
```

```
> cor(y,x^2)
```

```
[1] 0.9739695
```

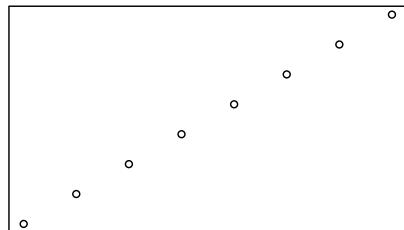
```
> x <- c(-5:5)
```

```
> cor(y,x^2)
```

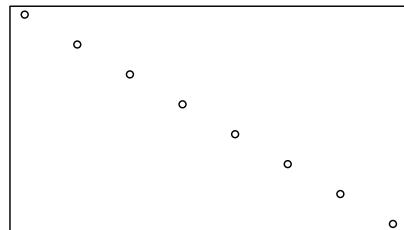
```
[1] 0
```

Correlation: Figure 9.11

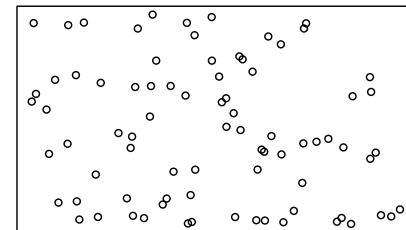
a. $r=1$



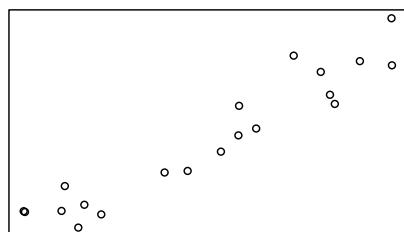
b. $r=-1$



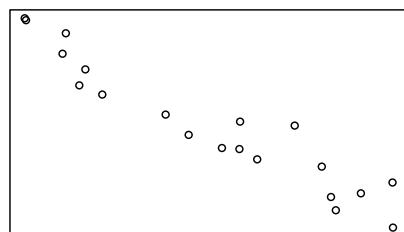
c. $r=0$



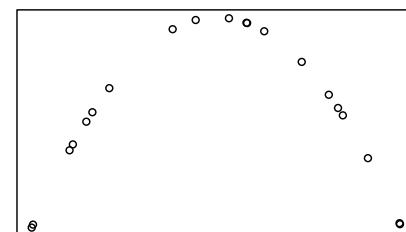
d. $0 < r < 1$



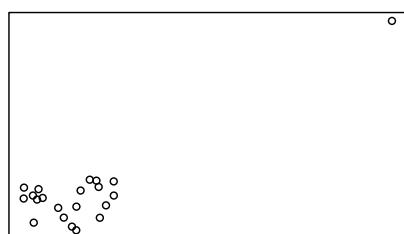
e. $-1 < r < 0$



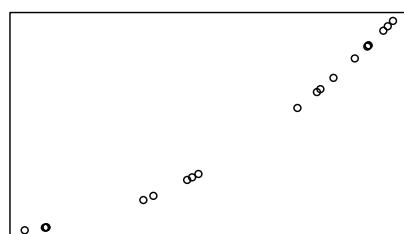
f. $r=0$



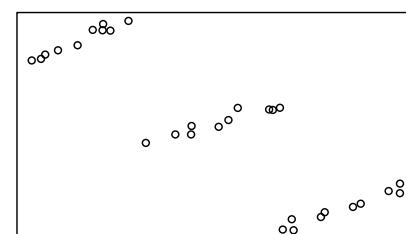
g. $0 < r < 1$



h. $0 < r < 1$



i. $-1 < r < 0$



Correlation

- The test statistic

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} \sim t_{N-2}$$

can be used to test $H_0 : \rho = 0$

- Claim: this test is equivalent to testing $H_0 : \beta_{y \cdot x} = 0$
- Proof of claim on next couple of pages

Correlation

- First note that

$$\begin{aligned}(N - 2)s_{y \cdot x}^2 &= \text{SSE} = \text{SST} - \text{SSR} \\ &= \text{SST} \left(1 - \frac{\text{SSR}}{\text{SST}} \right) \\ &= [Y^2] \left(1 - \frac{[XY]^2}{[Y^2][X^2]} \right) \\ &= (N - 1)s_Y^2(1 - r^2)\end{aligned}$$

- Next recall that

$$\hat{\beta}_{y \cdot x} = \frac{[XY]}{[X^2]}$$

Correlation

- Then

$$\begin{aligned} t &= \frac{\hat{\beta}_{y \cdot x}}{s_{y \cdot x} / \sqrt{[X^2]}} = \frac{[XY]/[X^2]}{s_{y \cdot x} / \sqrt{[X^2]}} \\ &= \frac{[XY]/\sqrt{[X^2]}}{s_{y \cdot x}} = \frac{r\sqrt{[Y^2]}}{s_{y \cdot x}} \\ &= \frac{rs_Y\sqrt{N-1}}{\sqrt{(1-r^2)s_Y^2(N-1)/(N-2)}} \\ &= \frac{r}{\sqrt{(1-r^2)/(N-2)}} \end{aligned}$$

Correlation

- In general,

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} \sim t_{N-2} \quad (1)$$

if

1. (X, Y) bivariate normal (Section 9.3.3 of the text), or
 2. $Y|X$ is normally distributed with constant variance
(that is, the usual regression model holds)
- (1) holds approximately for large N (cf. Graybill, 1976, Section 6.10)

Correlation Example

- Cholesterol was measured in 100 spouse pairs
- If there is no environmental effect (e.g., shared diet) on cholesterol we would expect $\rho = 0$
- $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$
- $t_{98,0.975} = 1.98$, so $C_{0.05} = \{t : |t| > 1.98\}$
- Observed $r = 0.25$, so that

$$t = \frac{r}{\sqrt{(1 - r^2)/(N - 2)}} = \frac{0.25}{\sqrt{(1 - 0.25^2)/98}} = 2.556$$

- $p = 2 \times \{1 - F_{t_{98}}(2.556)\} = 0.0121$

Correlation Example: SAS

```
proc corr;  
var x y;
```

Pearson Correlation Coefficients, N = 100

Prob > |r| under H0: Rho=0

	x	y
x	1.00000	0.25000
		0.0121
y	0.25000	1.00000
		0.0121

Correlation Using Fisher's Transformation

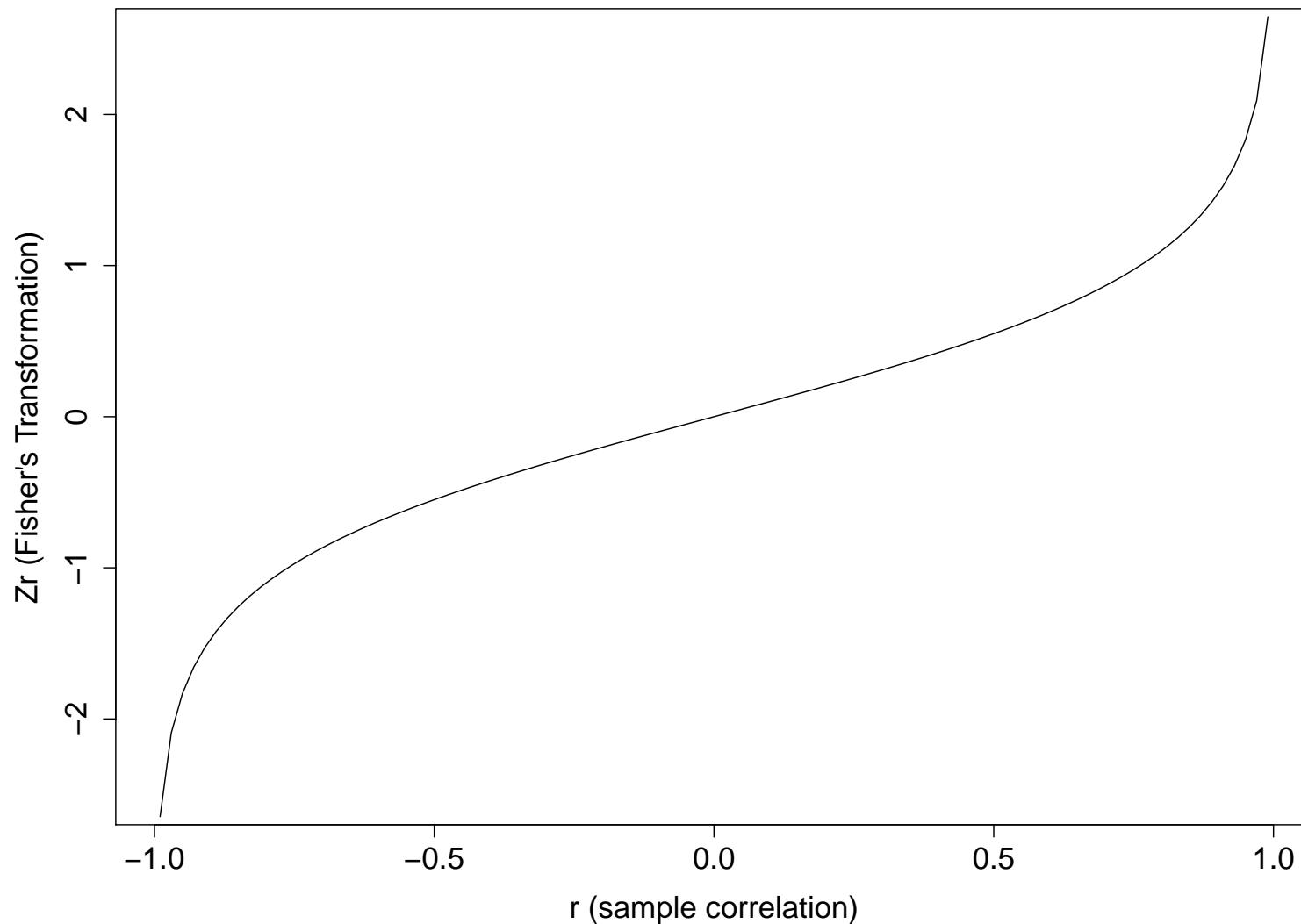
- R. A. Fisher developed a test of $H_0 : \rho = \rho_0$
- He showed that

$$z_r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \sim N \left(\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{N-3} \right)$$

- Under $H_0 : \rho = \rho_0$

$$z = \frac{\frac{1}{2} \log \left(\frac{1+r}{1-r} \right) - \frac{1}{2} \log \left(\frac{1+\rho_0}{1-\rho_0} \right)}{\sqrt{1/(N-3)}} \sim N(0, 1)$$

Correlation: Fisher's Transformation



Using Fisher's Transformation: Example

- Cholesterol example
- $N = 100, r = 0.25$
- $H_0 : \rho = 0$

$$z_r = \frac{1}{2} \log \left(\frac{1.25}{0.75} \right) = 0.2554$$

$$z = \frac{0.2554 - 0}{\sqrt{1/97}} = 2.5155$$

$$p = 2 \times \{1 - \Phi(2.515)\} = 0.0119$$

Correlation Using Fisher's Transformation

- The Fisher transformation can be used for a CI for ρ

$$z_r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \Rightarrow e^{2z_r} = \frac{1+r}{1-r} \Rightarrow r = \frac{e^{2z_r} - 1}{e^{2z_r} + 1}$$

$$z_L = z_r - z_{1-\alpha/2} \sqrt{1/(N-3)}$$

$$z_U = z_r + z_{1-\alpha/2} \sqrt{1/(N-3)}$$

$$r_L = \frac{e^{2z_L} - 1}{e^{2z_L} + 1}; \quad r_U = \frac{e^{2z_U} - 1}{e^{2z_U} + 1}$$

Using Fisher's Transformation: Example

- 95% CI when $r = 0.25$ and $n = 100$

$$(z_L, z_U) = 0.2554 \pm 1.96/\sqrt{97} = (0.0564, 0.4544)$$

$$r_L = \frac{e^{2 \times 0.0564} - 1}{e^{2 \times 0.0564} + 1} = 0.0563$$

$$r_U = \frac{e^{2 \times 0.4544} - 1}{e^{2 \times 0.4544} + 1} = 0.4255$$

Correlation Using Fisher's Transformation: SAS

```
proc corr fisher(biasadj=no);  
var x y;
```

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z
x	y	100	0.25000	0.25541

Pearson Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	95% Confidence Limits	p Value for $H_0: \text{Rho}=0$
x	y	0.056350	0.425524

Correlation Using Fisher's Transformation: R

```
> cor.test(x,y)
```

```
Pearson's product-moment correlation

data: x and y
t = 2.556, df = 98, p-value = 0.01212
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.05634962 0.42552363
sample estimates:
cor
0.2500007
```

Correlation Using Fisher's Transformation

- Comparing two correlations: Two independent samples

$$H_0 : \rho_1 = \rho_2 \text{ vs. } H_A : \rho_1 \neq \rho_2$$

- Compute z_{r_1} and z_{r_2}

$$\text{Var}(z_{r_1} - z_{r_2}) = \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}$$

- Thus under H_0

$$z = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \sim N(0, 1)$$

Using Fisher's Transformation: Example

- If blood pressure level is inherited, one would expect the correlation between blood pressure of mothers and their natural children to be greater than between mothers and their adopted children
- In a study, 1000 mothers and one of their randomly chosen natural children had their blood pressure measured
- In a separate sample, 100 mothers and their adopted children also had their BP measured

Using Fisher's Transformation: Example cont.

- Let

ρ_1 = population correlation for natural pairs

ρ_2 = population correlation for adopted pairs

- Hypotheses

$$H_0 : \rho_1 = \rho_2 \text{ vs. } H_A : \rho_1 > \rho_2$$

- Critical region

$$C_{0.05} = \{z : z > 1.645\}$$

Using Fisher's Transformation: Example cont.

- $r_1 = 0.32; r_2 = 0.06$
- $z_{r_1} = 0.3316; z_{r_2} = 0.0601$

- Thus

$$z = \frac{0.3316 - 0.0601}{\sqrt{\frac{1}{997} + \frac{1}{97}}} = 2.55$$

- So we reject the null hypothesis and conclude that blood pressure levels appear to have an inherited component

Correlation Homogeneity

- Testing the homogeneity of k correlations
- Fisher's transformation can be used to test the hypothesis that several correlations are equal

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_k$$

VS.

$$H_A : \text{at least one inequality}$$

Correlation Homogeneity

- Let

$$T_1 = \sum_{i=1}^k (n_i - 3) z_{r_i}$$

and

$$T_2 = \sum_{i=1}^k (n_i - 3) z_{r_i}^2$$

- Under H_0

$$H = T_2 - \frac{T_1^2}{\sum(n_i - 3)} \sim \chi_{k-1}^2$$

- Cf. Graybill (1976, p. 405)

Correlation Homogeneity: Example

- Does the correlation between LDL-cholesterol and HDL-cholesterol change with age in women not taking hormones?

Age	n	r	z_r
20-29	277	-0.08	-0.0802
30-39	479	-0.25	-0.2554
40-49	508	-0.19	-0.1923
50-59	373	-0.18	-0.1820
60-69	216	-0.15	-0.1511

Correlation Homogeneity: Example cont.

- Null hypothesis $H_0 : \rho_1 = \rho_2 = \cdots = \rho_k$
- Critical region $C_{0.05} = \{H : H > 9.49\}$
- Compute test statistic

$$T_1 = 274(-0.0802) + \cdots + 213(-0.1511) = -340.200$$

$$T_2 = 274(-0.0802)^2 + \cdots + 213(-0.1511)^2 = 68.614$$

$$H = 68.614 - \frac{(-340.200)^2}{1838} = 5.65$$

- So we do not reject the null hypothesis; the correlation between LDL-cholesterol and HDL-cholesterol does not appear to change with age

Rank Correlation Coefficients

- Using ranks makes statistics robust to outliers
- Spearman rank correlation, Kendall's τ
- Nonparametric measures of association

Spearman Rank Correlation

1. Y_s and X_s are ranked from 1 to N separately
2. The correlation of the ranks is then computed

Spearman Correlation: Example

- Ten children are ranked according to their mathematical and musical abilities

Child	Math	Music
A	7	5
B	4	7
C	3	3
D	10	10
E	6	1
F	2	9
G	9	6
H	8	2
I	1	8
J	5	4

Spearman Correlation

- Let R_{1i} and R_{2i} be the ranks of the Y_i and X_i , respectively
- Spearman correlation coefficient

$$\begin{aligned} r_s &= \frac{\sum(R_{1i} - \bar{R}_1)(R_{2i} - \bar{R}_2)}{\sqrt{\sum_i(R_{1i} - \bar{R}_1)^2 \sum_i(R_{2i} - \bar{R}_2)^2}} \\ &= 1 - \frac{6 \sum d_i^2}{N^3 - N} \end{aligned}$$

where $d_i = R_{1i} - R_{2i}$

- The form of r_s containing $\sum d_i^2$ is not correct if ties are present
- Note:

$$R_{1i} = R_{2i} \text{ for all } i \Rightarrow d_i = 0 \text{ for all } i \Rightarrow r_s = 1$$

Spearman Correlation

- Suppose N is odd and $N = 2m + 1$
- Then the most extreme discordant rankings are

i	1	2	\dots						N
R_{1i}	1	2	\dots	m	$m + 1$	$m + 2$	\dots	$2m$	$2m + 1$
R_{2i}	$2m + 1$	$2m$	\dots	$m + 2$	$m + 1$	m	\dots	2	1
d_i	$-2m$	$2 - 2m$	\dots	-2	0	2	\dots	$2m - 2$	$2m$

Spearman Correlation

- Under this configuration

$$\begin{aligned}\sum_{i=1}^N d_i^2 &= 4m^2 + 4(m-1)^2 + \cdots + 4(1)^2 + 0 \\ &\quad + 4(1)^2 + \cdots + 4(m-1)^2 + 4m^2 \\ &= 8 \sum_{j=1}^m j^2 \\ &= 8m(m+1)(2m+1)/6 \\ &= \left(4 \times \frac{N-1}{2} \times \frac{N+1}{2} \times N\right) / 3 \\ &= (N^3 - N)/3\end{aligned}$$

Spearman Correlation

- Thus

$$r_s = 1 - \frac{6(N^3 - N)}{3(N^3 - N)} = 1 - 2 = -1$$

- In a similar way, it can be shown that if N is even, the most extreme rankings give $r_s = -1$
- So:

$r_s = 1$ if perfect agreement in the ranks

$r_s = -1$ if perfect disagreement in the ranks

Spearman Correlation

	Child	Math	Music	d
A	7	5	2	
B	4	7	-3	
C	3	3	0	
D	10	10	0	
E	6	1	5	
F	2	9	-7	
G	9	6	3	
H	8	2	6	
I	1	8	-7	
J	5	4	1	

- Spearman correlation

$$r_s = 1 - \frac{6(2^2 + (-3)^2 + \dots + 1^2)}{10^3 - 10} = 1 - \frac{6(182)}{990} = -0.103$$

Spearman Correlation: SAS and R

```
proc corr spearman;  
  var math music;
```

Spearman Correlation Coefficients, N = 10

Prob > |r| under H0: Rho=0

	math	music
math	1.00000	-0.10303
		0.7770
music	-0.10303	1.00000
		0.7770

```
> cor(math,music,method="spearman")  
[1] -0.1030303
```

Spearman Correlation

- The Spearman correlation coefficient can be used to test the null hypothesis of independence

$$H_0 : X \perp Y \text{ vs. } H_A : X \not\perp Y$$

that is, H_A : X and Y not independent

- Distribution of r_s under H_0 is derived using a permutation-based argument
- We can list the R_{1i} in ascending order
- There are $N!$ possible orderings of the R_{2i}
- Under H_0 , each of these orderings is equally likely

Spearman Correlation

- Example: $N = 3$

R_{1i}	1	2	3	$\sum d_i^2$	r_s
R_{2i}	1	2	3	0	1.0
R_{2i}	1	3	2	2	0.5
R_{2i}	2	1	3	2	0.5
R_{2i}	2	3	1	6	-0.5
R_{2i}	3	1	2	6	-0.5
R_{2i}	3	2	1	8	-1.0

Spearman Correlation

- CDF of r_s

k	$\Pr[r_s \leq k]$
-1.0	1/6
-0.5	1/2
0.5	5/6
1.0	1

- Text Table A.12, p. 838, gives the two sided critical values for testing $H_0 : X \perp Y$
- If N is large (> 10 ; Neter et al. 1996, page 652),

$$t_s = \frac{r_s \sqrt{N - 2}}{\sqrt{1 - r_s^2}} \sim t_{N-2}$$

Spearman Correlation: Example

- Example: math (X) and music (Y)
- $N = 10$; $r_s = -0.1030$
- From Table A.12, $C_{0.05} = \{r_s : |r_s| > 0.648\}$
- Assume $N = 10$ is large enough to use the t approximation
- $C_{0.05} = \{t_s : |t_s| > t_{8,0.975} = 2.306\}$
- $t_s = \frac{-0.1030\sqrt{8}}{\sqrt{1-(-0.1030)^2}} = -0.2930$
- $p = 2 \times \Pr[t_8 < -0.2929] = 0.7771$

Spearman Correlation: Ties

- In the presence of ties, ranks are replaced by midranks
- However, critical values in Table A.12 are only approximate
- If N is large, use t_s as before; i.e.,

$$t_s = \frac{r_s \sqrt{N - 2}}{\sqrt{1 - r_s^2}} \sim t_{N-2}$$

Kendall's τ

- Kendall's τ : Another rank correlation statistic
- Data: (X_i, Y_i) for $i = 1, 2, \dots, N$
- Definitions: Two pairs of observations are
 - concordant if $(X_i - X_j)(Y_i - Y_j) > 0$
 - discordant if $(X_i - X_j)(Y_i - Y_j) < 0$

Kendall's τ

- Let p_c be the probability that a randomly chosen pair of observations is concordant; and p_d the probability that they are discordant; then

$$\tau = p_c - p_d$$

- Note:

$$-1 \leq \tau \leq 1$$

if X and Y are independent, $\tau = 0$

Kendall's τ

- There are $\binom{N}{2}$ pairs of observations
- Let P be the number of concordant pairs
- Let Q be the number of discordant pairs
- The estimate of τ is

$$r_k = \frac{P - Q}{\binom{N}{2}} = 1 - \frac{2Q}{\binom{N}{2}} = \frac{2P}{\binom{N}{2}} - 1$$

- The last two terms assume no ties, so that $P + Q = \binom{N}{2}$
- Replacing X s and Y s with their ranks does not change τ

Kendall's τ

- $H_0 : \tau = 0$ vs. $H_A : \tau \neq 0$
- The distribution of r_k under H_0 is computed using permutation principles
- As with r_s , there are $N!$ equally likely outcomes
- Kendall, *Rank Correlation Methods*, Hafner Publishing, 1962, gives a table of the distribution of $P - Q$ for $4 \leq N \leq 10$

Kendall's τ

- Upper one-sided critical values of r_k
- Note that the distribution of r_k is symmetric about 0

N	0.05	0.025
5	0.80	1.00
6	0.73	0.87
7	0.62	0.71
8	0.57	0.64
9	0.50	0.56
10	0.42	0.51

Kendall's τ : Example

- Cigarette consumption and lung cancer mortality in England and Wales, 1930-1969

Period	\log_{10} mortality	\log_{10} tobacco (lb/person)
1930-34	-2.35	-0.26
1935-39	-2.20	-0.03
1940-44	-2.12	0.30
1945-49	-1.95	0.37
1950-54	-1.85	0.40
1955-59	-1.80	0.50
1960-64	-1.70	0.55
1965-69	-1.58	0.55

Kendall's τ

- $C_{0.05} = \{r_k : |r_k| \geq 0.64\}$
- Observation 1: $(-2.35, -0.26)$
Observation 2: $(-2.20, -0.03)$
 $\{-2.35 - (-2.2)\}\{-0.26 - (-0.03)\} > 0 \Rightarrow$ concordant
- Observation 1 and observation 3:
 $\{-2.35 - (-2.12)\}(-0.26 - 0.3) > 0 \Rightarrow$ concordant
- $P - Q = 27 \Rightarrow$

$$r_k = \frac{27}{\binom{8}{2}} = \frac{27}{28} = 0.96$$

Kendall's τ

- If N is sufficiently large (≥ 10), under $H_0 : \tau = 0$

$$r_k \sim N\left(0, \frac{2(2N+5)}{9N(N-1)}\right)$$

$$P - Q \sim N\left(0, \frac{N(N-1)(2N+5)}{18}\right)$$

or

$$Z = \frac{P - Q}{\sqrt{\frac{N(N-1)(2N+5)}{18}}} \sim N(0, 1)$$

Kendall's τ

- If there are tied observations, r_k cannot be 1 or -1 .
- Let

$$t_x = \frac{1}{2} \sum_i t_{xi}(t_{xi} - 1) \quad \text{and} \quad t_y = \frac{1}{2} \sum_i t_{yi}(t_{yi} - 1)$$

where t_{zi} denotes the number of observations in the i^{th} set of ties for $z = x, y$

Kendall's τ

- Let

$$W = \sqrt{\left(\frac{1}{2}N(N-1) - t_x\right)\left(\frac{1}{2}N(N-1) - t_y\right)}$$

- Define

$$r_{k_b} = \frac{P - Q}{W}$$

This statistic is known as *Kendall's τ_b*

Kendall's τ : Tobacco Example Revisited

- Recall that $N = 8$ and there was one set of ties (of size 2) for the tobacco variable
- Thus

$$W = \sqrt{\left(\frac{1}{2}8(8 - 1)\right)\left(\frac{1}{2}8(8 - 1) - 1\right)}$$

- Yielding

$$r_{k_b} = \frac{27}{\sqrt{28 \times 27}} = 0.98198$$

Kendall's τ : Tobacco Example cont.

- SAS

```
proc corr kendall;  
var mortality tobacco;
```

Kendall Tau b Correlation Coefficients, N = 8
Prob > |tau| under H0: Tau=0

	mortality	tobacco
mortality	1.00000	0.98198 0.0008
tobacco	0.98198 0.0008	1.00000

Kendall's τ : Tobacco Example cont.

- R

```
> cor(mortality, tobacco, method="kendall")
[1] 0.9819805
```

```
> cor.test(mortality, tobacco, method="kendall")
```

Kendall's rank correlation tau

```
data: mortality and tobacco
z = 3.3662, p-value = 0.000762
alternative hypothesis: true tau is not equal to 0
sample estimates:
```

```
tau
0.9819805
```

Warning message:

```
In cor.test.default(mortality, tobacco, method = "kendall") :
  Cannot compute exact p-value with ties
```

Kendall's τ

- Kendall's score $P - Q$

$$r_{k_a} = \frac{P - Q}{\binom{N}{2}}$$

and

$$r_{k_b} = \frac{P - Q}{W}$$

- Tests based on r_{k_a} and r_{k_b} are equivalent
- Asymptotic variance of $P - Q$ under H_0 is given on page 336 of the text

$$Z = \frac{P - Q}{\sqrt{\text{Var}(P - Q)}} \sim N(0, 1)$$

Kendall's τ : Example

- In general, $\text{Var}(P - Q)$ equals

$$\frac{N(N-1)(2N+5)}{18} - \sum_i \frac{t_{xi}(t_{xi}-1)(2t_{xi}+5)}{18} - \dots$$

- For tobacco example, $\text{Var}(P - Q)$ is

$$\frac{8(8-1)(2 \cdot 8 + 5)}{18} - 0 - \frac{2(2-1)(2 \cdot 2 + 5)}{18} + 0 + 0 = 64.333$$

- Thus

$$z = \frac{27}{\sqrt{64.333}} = 3.366$$

yielding $p = 2 \cdot \{1 - \Phi(3.366)\} = 0.0008$

Correlation: Summary/Remarks

- r is appropriate if (X, Y) bivariate normal; sensitive to outliers, major(?) departures from normality
- Nonparametric alternatives: r_s and r_k
- If (X, Y) bivariate normal with correlation ρ ,

$$r \xrightarrow{p} \rho \quad r_s \xrightarrow{p} \frac{6}{\pi} \arcsin(\rho/2) \quad r_k \xrightarrow{p} \frac{2}{\pi} \arcsin(\rho)$$

(Kraemer, 1998 “Rank Correlation” *Encyclopedia of Biostatistics*)

- ARE of r_s and r_k compared to r : $9/\pi^2 = 0.912$
(Conover, 1980 *Practical Nonparametric Statistics*)

BIOS 662 Fall 2018

Analysis of Variance, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Introduction
- Alternative models
- SS decomposition
- Example using SAS, R

Analysis of Variance Model

- Chapter 10 of the text (skip 10.3-10.5); chapter 12
- How do we test hypotheses about the mean of more than two groups? Analysis of variance (ANOVA) model
- *Definition 10.1:* An *analysis of variance model* is a linear regression model in which the predictor variables are classification variables. The categories of a variable are called the *levels* of the variable.
- Categorical predictor variables are also called *qualitative factors*

Notation

- Let Y_{ij} be the j^{th} observation in the i^{th} group
- $i = 1, \dots, K; j = 1, \dots, n_i$
- Let $N = \sum_{i=1}^K n_i$
- $\bar{Y}_{i\cdot} = \sum_j Y_{ij} / n_i$

ANOVA Model and Hypotheses

- Assume $Y_{ij} \sim N(\mu_i, \sigma^2)$

- Want to test

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$$

versus

$$H_A : \text{at least one inequality}$$

Two Variance Estimators

- The pooled estimator of σ^2 is:

$$s_p^2 = \frac{\sum_{i=1}^K (n_i - 1) s_i^2}{\sum_{i=1}^K (n_i - 1)}$$

- Under H_0 , the (weighted) variance of the $\bar{Y}_{i\cdot}$ s will estimate σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^K n_i (\bar{Y}_{i\cdot} - \bar{Y})^2}{K - 1}$$

where

$$\bar{Y} = \frac{\sum_{i=1}^K \sum_{j=1}^{n_i} Y_{ij}}{N}$$

ANOVA: F Test

- It can be shown that under H_0 :

$$(N - K)s_p^2/\sigma^2 \sim \chi_{N-K}^2$$

$$(K - 1)\hat{\sigma}^2/\sigma^2 \sim \chi_{K-1}^2$$

and s_p^2 and $\hat{\sigma}^2$ are independent

- Therefore, under H_0 ,

$$F \equiv \frac{\hat{\sigma}^2}{s_p^2} \sim F_{K-1, N-K}$$

ANOVA: F Test

- To test H_0 ,

$$C_\alpha = \{F : F > F_{K-1, N-K; 1-\alpha}\}$$

- The test uses $F > F_{K-1, N-K; 1-\alpha}$ because under H_A ,

$$E(\hat{\sigma}^2) > E(s_p^2)$$

- In particular, $E(s_p^2) = \sigma^2$, whereas

$$E(\hat{\sigma}^2) = \sigma^2 + \frac{\sum_i n_i (\mu_i - \mu)^2}{K - 1}$$

where μ is the overall mean defined in equation (1) a few pages ahead

ANOVA: Example

- Passive smoking and lung function
- A study was conducted to compare the lung function of groups of smokers and non-smokers. Lung function was measured by forced expiratory flow (FEF)
- FEF for males by smoking status:

Group	n_i	Mean (L/sec)	sd (L/sec)
Non-smokers	200	3.78	0.79
Passive smokers	200	3.30	0.77
Non-inhalers	50	3.32	0.86
Light smokers	200	3.23	0.78
Mod. smokers	200	2.73	0.81
Heavy smokers	200	2.59	0.82

ANOVA: Example cont.

$$C_{0.05} = \{F > F_{5,1044;0.95} = 2.22\}$$

$$s_p^2 = \frac{199(0.79)^2 + 199(0.77)^2 + \cdots + 199(0.82)^2}{1044} = 0.636$$

$$\hat{\sigma}^2 = \frac{200(3.78 - 3.158)^2 + \cdots + 200(2.59 - 3.158)^2}{5} = 36.987$$

- $F = 36.987/0.636 = 58.17 > 2.22$; so reject H_0
- Reference: White JR, Froeb HF. *N Engl J Med* 302(13): 720-3, 1980. (Results presented here may differ from those in the original manuscript because of rounding.)

Aside: Obtaining Quantiles/CDFs

- In R

```
> qf(0.95,5,1044)
```

```
[1] 2.222674
```

```
> 1-pf(58.17,5,1044)
```

```
[1] 0
```

- In SAS

```
data;  
    y = finv(0.95,5,1044);  
    y1 = quantile('F',0.95,5,1044);  
    fy = cdf('F',58.17,5,1044);  
  
proc print;
```

Obs	y	y1	fy
1	2.22267	2.22267	1

Cell Means Model

- The version of the ANOVA model we have looked at so far is called the *cell means model*

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, 2, \dots, K; j = 1, 2, \dots, n_i$ where

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

Factor Effects Model

- An equivalent model is the *factor effects model*

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

for $i = 1, 2, \dots, K; j = 1, 2, \dots, n_i$ where

$$\mu = \frac{1}{N} \sum_{i=1}^K n_i \mu_i \quad (1)$$

$$\alpha_i = \mu_i - \mu$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

- Note typo in the text on page 363
- Here α_i does not denote type I error

Factor Effects Model

- Constraint: $\sum_{i=1}^K n_i \alpha_i = 0$
- Suppose $K = 4$, then from the constraint,

$$n_1 \alpha_1 + n_2 \alpha_2 + n_3 \alpha_3 + n_4 \alpha_4 = 0$$

and so

$$\alpha_4 = -(n_1 \alpha_1 + n_2 \alpha_2 + n_3 \alpha_3) / n_4$$

Thus

$$Y_{1j} = \mu + 1\alpha_1 + \epsilon_{1j}$$

$$Y_{2j} = \mu + 1\alpha_2 + \epsilon_{2j}$$

$$Y_{3j} = \mu + 1\alpha_3 + \epsilon_{3j}$$

$$Y_{4j} = \mu - \frac{n_1}{n_4} \alpha_1 - \frac{n_2}{n_4} \alpha_2 - \frac{n_3}{n_4} \alpha_3 + \epsilon_{4j}$$

Model Equivalence

- Equivalence of null hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \Leftrightarrow H_0 : \alpha_i = 0; \quad i = 1, 2, \dots, K$$

- α_i is called the i^{th} *main effect* or *factor effect*

$$\begin{aligned} Y_{ij} &= \mu_i + \epsilon_{ij} \\ &= \mu + (\mu_i - \mu) + \epsilon_{ij} \\ &= \mu + \alpha_i + \epsilon_{ij} \\ &= \text{mean} + i^{\text{th}} \text{ main effect} + \text{error} \end{aligned}$$

- Data can be partitioned similarly

$$\begin{aligned} Y_{ij} &= \bar{Y} + (\bar{Y}_{i\cdot} - \bar{Y}) + (Y_{ij} - \bar{Y}_{i\cdot}) \\ &= \bar{Y} + a_i + e_{ij} \end{aligned}$$

Reference Group Model

- Another equivalent model is the *reference group model*
- One group is chosen as the reference; suppose it is group 1
- Then

$$Y_{1j} = \mu_1 + \epsilon_{1j}$$

$$\begin{aligned} Y_{ij} &= \mu_1 + (\mu_i - \mu_1) + \epsilon_{ij}, \quad i = 2, 3, \dots, K \\ &= \mu_1 + \beta_i + \epsilon_{ij}, \quad i = 2, 3, \dots, K \end{aligned}$$

for

$$j = 1, 2, \dots, n_i$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

- Null hypothesis:

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_K = 0$$

ANOVA: Sum of Squares

- It can be shown (see a few pages ahead) that

$$\sum_{i=1}^K \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \bar{Y})^2 + \sum_{i=1}^K \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2$$

- That is,

$$SST = SSA + SSW$$

$$= (K - 1)\hat{\sigma}^2 + (N - K)s_p^2$$

- SSW is also referred to as SSE

ANOVA: Sum of Squares

- Expected value of sum of squares

$$E\left(\sum_{i=1}^K n_i(\bar{Y}_{i\cdot} - \bar{Y})^2\right) = \sum_{i=1}^K n_i\alpha_i^2 + (K-1)\sigma^2$$

$$E\left(\sum_{i=1}^K \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2\right) = (N-K)\sigma^2$$

- Under $H_0 : \alpha_1 = \dots = \alpha_K = 0$,

$$E\left(\sum_{i=1}^K n_i(\bar{Y}_{i\cdot} - \bar{Y})^2\right) = (K-1)\sigma^2$$

ANOVA: F Test and ANOVA Table

- Therefore, under H_A : at least one $\alpha_i \neq 0$,

$$E(F) > 1$$

- That is, we reject H_0 if F is too large

$$C_\alpha = \{F : F > F_{K-1, N-k; 1-\alpha}\}$$

ANOVA Table

Source of variation	df	MS	F
Among groups	$K - 1$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^K n_i (\bar{Y}_{i\cdot} - \bar{Y})^2}{K-1}$	MSA/MSW
Within groups	$N - K$	$s_p^2 = \frac{\sum_{i=1}^K (n_i - 1)s_i^2}{N-K}$	
Total	$N - 1$		

ANOVA: Sum of Squares Proof

- Start with

$$\sum_{ij} (Y_{ij} - \bar{Y})^2 = \sum_{ij} (Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y})^2$$

- The RHS is equivalent to

$$\sum_{ij} (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_{ij} (\bar{Y}_{i\cdot} - \bar{Y})^2 + 2 \sum_{ij} (Y_{ij} - \bar{Y}_{i\cdot})(\bar{Y}_{i\cdot} - \bar{Y})$$

- The last term can be written as

$$2 \sum_i \left((\bar{Y}_{i\cdot} - \bar{Y}) \sum_j (Y_{ij} - \bar{Y}_{i\cdot}) \right)$$

which equals zero because

$$\sum_j (Y_{ij} - \bar{Y}_{i\cdot}) = 0 \quad \text{for all } i$$

ANOVA: $E(\text{SSW})$ Proof

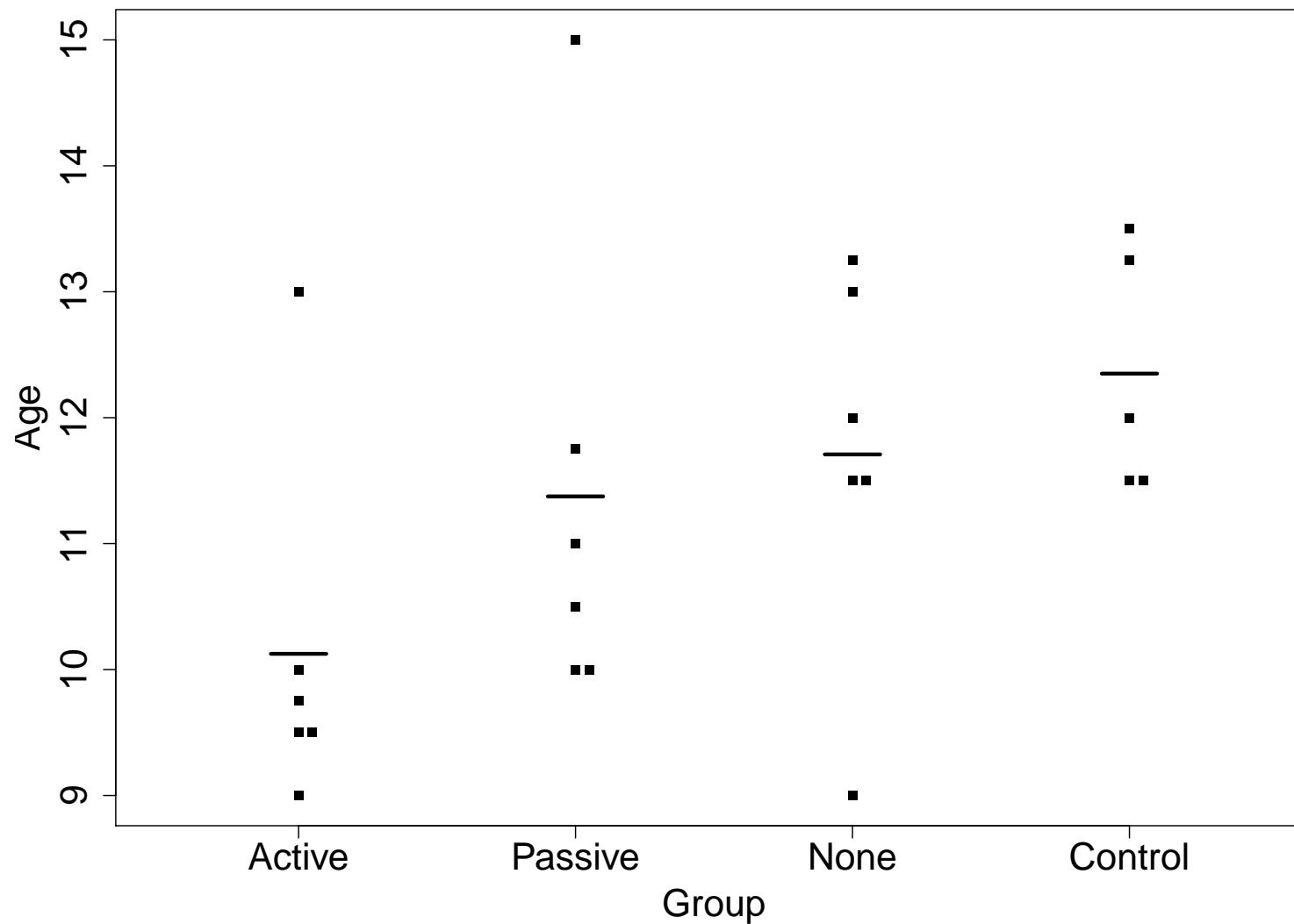
$$\begin{aligned} E(\text{SSW}) &= E\left(\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2\right) \\ &= E\left(\sum_i(n_i - 1)\frac{\sum_j(Y_{ij} - \bar{Y}_{i\cdot})^2}{n_i - 1}\right) \\ &= \sum_i(n_i - 1)E(s_i^2) \\ &= \sum_i(n_i - 1)\sigma^2 \\ &= (N - K)\sigma^2 \end{aligned}$$

ANOVA: Example

- Table 10.1: Distribution of ages (in months) at which infants first walked alone

Active Group	Passive Group	No-Exercise Group	Eight-week Control group
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	

ANOVA: Example cont.



ANOVA: SAS – Cell Means Model

```
proc anova data=one;
* Using the following proc statement yields exactly the same ANOVA table;
* proc glm data=one;
  class group;
  model age=group;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14.77780797	4.92593599	2.14	0.1285
Error	19	43.68958333	2.29945175		
Corrected Total	22	58.46739130			

ANOVA: SAS – Factor Effects Model

```
data two;  
  set one;  
  x1=0; x2=0; x3=0;  
  if group="active" then x1=1;  
  else if group="passive" then x2=1;  
  else if group="no" then x3=1;  
  else if group="eight" then do; x1=x2=x3=-6/5; end;  
  
proc reg data=two;  
  model age = x1  x2  x3;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14.77781	4.92594	2.14	0.1285
Error	19	43.68958	2.29945		
Corrected Total	22	58.46739			

ANOVA: SAS – Reference Group Model

```
data three;
  set one;
  x2=0; x3=0; x4=0;
  if group="passive" then x2=1;
  else if group="no" then x3=1;
  else if group="eight" then x4=1;

proc reg data=three;
  model age = x2  x3  x4;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	14.77781	4.92594	2.14	0.1285
Error	19	43.68958	2.29945		
Corrected Total	22	58.46739			

ANOVA: R

```
> group <- as.factor(group)
> av <- aov(age ~ group)
> anova(av)
```

Analysis of Variance Table

Response: age

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	3	14.778	4.9259	2.1422	0.1285
Residuals	19	43.690	2.2995		

BIOS 662 Fall 2018

Analysis of Variance, Part II

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Multiple Comparisons
 - Scheffé
 - Tukey
 - Bonferroni
- Chapter 12 of the text

Multiple Comparisons

- Suppose we do n independent tests, each with probability α of making a type I error
- Suppose all n null hypotheses are true
- What is the probability of making at least one type I error?

$$1 - (1 - \alpha)^n$$

Multiple Comparisons

- Table 12.1: Probability of rejecting at least one null hypothesis when n independent tests are carried out at the α level and each null hypothesis is true

n	α		
	0.01	0.05	0.10
1	0.01	0.05	0.10
2	0.02	0.10	0.19
3	0.03	0.14	0.27
4	0.04	0.19	0.34
5	0.05	0.23	0.41
10	0.10	0.40	0.65
20	0.18	0.64	0.88
100	0.63	0.99	1.00

Multiple Comparisons

- Definition 12.2: The probability of incorrectly rejecting at least one of the true null hypotheses in an experiment involving one or more tests or comparisons is called the *per experiment error rate (PEER)*
- PEER is also known as the *family-wise error rate (FWE)*

ANOVA and Multiple Comparisons

- Rejection of $H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$ does not indicate where the inequalities are
- For example,

$$H_A : \mu_1 = \mu_2 = \cdots = \mu_{K-1} \neq \mu_K$$

or

$$H_A : \mu_1 \neq \mu_2 \neq \cdots \neq \mu_{K-1} \neq \mu_K$$

- Usually we want to identify the inequalities

ANOVA

- Need a multiple comparisons method to test the $\binom{K}{2}$ null hypotheses

$$H_0 : \mu_i = \mu_j \quad (i \neq j)$$

- Popular methods:
 - Scheffé
 - Tukey
 - Bonferroni (Sidak, Holm, Hochberg)

ANOVA: Scheffé

- For each pair of means, compute

$$t_{ij} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- Rejection region

$$C_\alpha = \left\{ t_{ij} : |t_{ij}| > \sqrt{(K-1)F_{K-1, N-K, 1-\alpha}} \right\}$$

- Passive smoking example

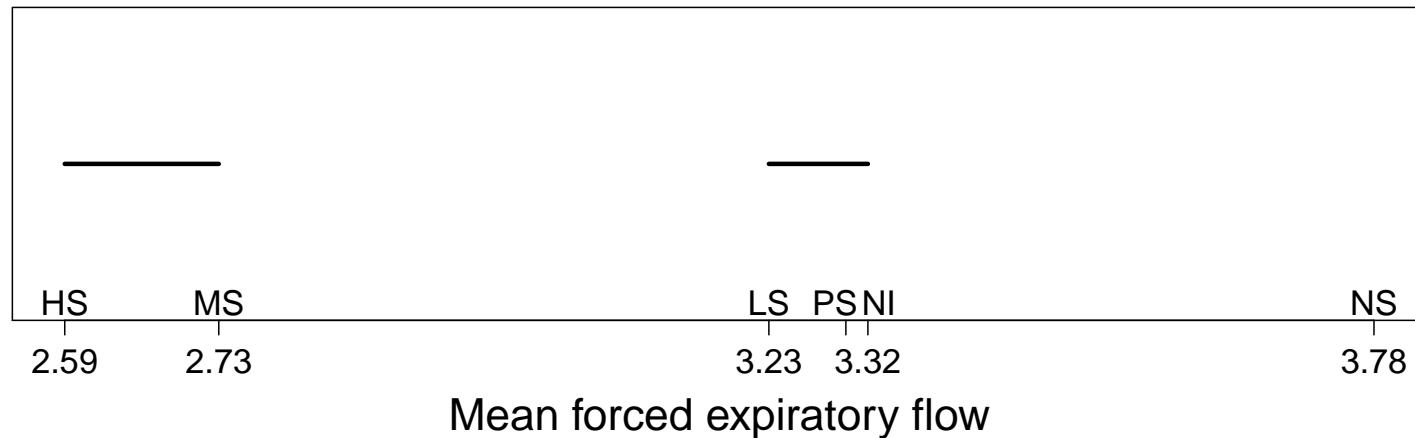
$$C_{0.05} = \left\{ t_{ij} : |t_{ij}| > \sqrt{5F_{5, 1044, 0.95}} = \sqrt{5(2.22)} = 3.33 \right\}$$

Scheffé: Passive Smoking Example

Comparison	t_{ij}	Significant
NS-PS	6.02	yes
NS-NI	3.65	yes
NS-LS	6.90	yes
NS-MS	13.17	yes
NS-HS	14.92	yes
PS-NI	-0.16	no
PS-LS	0.88	no
PS-MS	7.15	yes
PS-HS	8.90	yes
NI-LS	0.71	no
NI-MS	4.68	yes
NI-HS	5.79	yes
LS-MS	6.27	yes
LS-HS	8.03	yes
MS-HS	1.76	no

Scheffé: Passive Smoking Example cont.

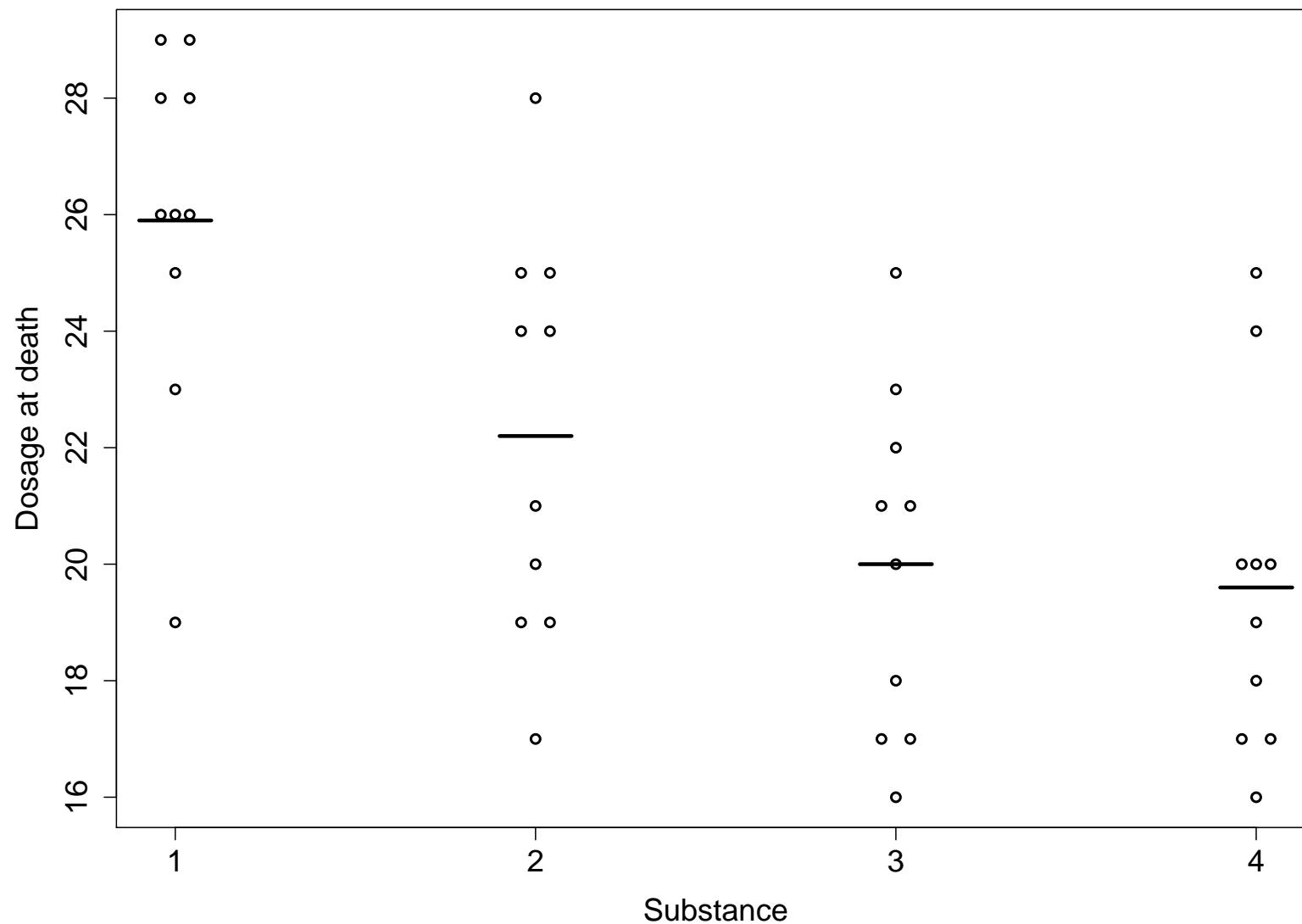
- Overall conclusions about similarities and differences across the population means indicated using schematic diagram
- Use overbars to connect means that do not differ significantly



Scheffé: Example II

- Four cardiac substances tested for relative potencies
- For each substance, ten guinea pigs anesthetized
- Outcome: dosage at death
- Data display with group means on following page

Scheffé: Example II cont.



Scheffé: Example II cont.

- Global F-test strongly rejects the null of equality of the four population means ($p = 0.0002$)
- Critical region

$$C_{0.05} = \left\{ t_{ij} : |t_{ij}| > \sqrt{3F_{3,36,0.95}} = \sqrt{3 \times 2.866} = 2.93 \right\}$$

- Note that in this example the denominator of t_{ij} is always $\sqrt{\text{MSE}/5} = 1.396$ because all group sizes are 10
- So we could also write the critical region in terms of the *minimum significant difference*

$$C_{0.05} = \{ |\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > 2.93 \times 1.396 = 4.09 \}$$

Scheffé: SAS

```
proc glm; class group; model dose=group; means group/scheffe;  
(proc anova with the same statements yields the same output as below)
```

Scheffe's Test for dose

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	36
Error Mean Square	9.747222
Critical Value of F	2.86627
Minimum Significant Difference	4.0942

Means with the same letter are not significantly different.

Scheffe Grouping		Mean	N	group
A		25.900	10	1
	A			
B	A	22.200	10	2
	B			
	B	20.000	10	3
	B			
	B	19.600	10	4

ANOVA: Scheffé

- For each pair of means, we can also compute multiplicity adjusted confidence intervals using Scheffé's method

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm \sqrt{\text{MSE} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \times \sqrt{(K - 1)F_{K-1, N-K, 1-\alpha}}$$

- The probability is at least $1 - \alpha$ that these intervals simultaneously straddle the corresponding population mean differences
- What happens when $K = 2$?
- For the cardiac substance example,

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm 4.09$$

Scheffé: SAS

```
proc glm; class group; model dose=group; means group/scheffe cldiff;  
(proc anova with the same statements yields the same output as below)
```

Scheffe's Test for dose

Comparisons significant at the 0.05 level are indicated by ***.

group	Comparison	Difference	Simultaneous	
		Between Means	95% Confidence Limits	
	1 - 2	3.700	-0.394	7.794
	1 - 3	5.900	1.806	9.994 ***
	1 - 4	6.300	2.206	10.394 ***
	2 - 1	-3.700	-7.794	0.394
	2 - 3	2.200	-1.894	6.294
	2 - 4	2.600	-1.494	6.694
	3 - 1	-5.900	-9.994	-1.806 ***
	3 - 2	-2.200	-6.294	1.894
	3 - 4	0.400	-3.694	4.494
	4 - 1	-6.300	-10.394	-2.206 ***
	4 - 2	-2.600	-6.694	1.494
	4 - 3	-0.400	-4.494	3.694

ANOVA: Tukey

- Alternative multiple comparisons approach to Scheffé
- Critical region

$$C_\alpha = \left\{ t_{ij} : |t_{ij}| > (q_{K,N-K,1-\alpha})/\sqrt{2} \right\}$$

where $q_{k,m,1-\alpha}$ is the $1 - \alpha$ quantile of the *studentized range*; see `qtukey` in R and `probmc('Range', ...)` in SAS

- Multiplicity adjusted CIs

$$\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} \pm \sqrt{\text{MSE} \times 2/n} \times (q_{K,N-K,1-\alpha})/\sqrt{2}$$

Note that the multiplicity adjusted CIs here assume a balanced design, that is, $n_i = n$ for all i

ANOVA: Tukey

- What is the studentized range?
- Suppose Y_1, \dots, Y_k iid $N(\mu, \sigma^2)$
- Let s be an estimator for σ with m degrees of freedom, $s \perp Y_1, \dots, Y_k$
- Then

$$\frac{Y_{(k)} - Y_{(1)}}{s}$$

has a studentized range distribution with parameters k and m

ANOVA: Tukey

- Cardiac substance example with $\alpha = 0.05$

$$q_{K,N-K,1-\alpha}/\sqrt{2} = q_{4,36,0.95}/\sqrt{2} = 2.69$$

- Compared with the Scheffé critical value (2.93), easier to reject; equivalently, Tukey confidence intervals will be narrower
- For this reason, Tukey is preferred to Scheffé in balanced designs where all pairwise comparisons are being considered
- Otherwise, use Scheffé or Bonferroni-type method (later in this section)

Tukey: SAS

```
proc glm; class group; model dose=group; means group/tukey cldiff;  
(proc anova with the same statements yields the same output as below)
```

Tukey's Studentized Range (HSD) Test for dose

Comparisons significant at the 0.05 level are indicated by ***.

group	Comparison	Difference	Simultaneous	
		Between Means	95% Confidence Limits	
	1 - 2	3.700	-0.060	7.460
	1 - 3	5.900	2.140	9.660 ***
	1 - 4	6.300	2.540	10.060 ***
	2 - 1	-3.700	-7.460	0.060
	2 - 3	2.200	-1.560	5.960
	2 - 4	2.600	-1.160	6.360
	3 - 1	-5.900	-9.660	-2.140 ***
	3 - 2	-2.200	-5.960	1.560
	3 - 4	0.400	-3.360	4.160
	4 - 1	-6.300	-10.060	-2.540 ***
	4 - 2	-2.600	-6.360	1.160
	4 - 3	-0.400	-4.160	3.360

Tukey: R

```
> group <- as.factor(group)
> fit <- aov(dose ~ group)
> TukeyHSD(fit,"group")
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = dose ~ group)

\$group

	diff	lwr	upr	p adj
2-1	-3.7	-7.460351	0.06035128	0.0551754
3-1	-5.9	-9.660351	-2.13964872	0.0008587
4-1	-6.3	-10.060351	-2.53964872	0.0003701
3-2	-2.2	-5.960351	1.56035128	0.4048758
4-2	-2.6	-6.360351	1.16035128	0.2621133
4-3	-0.4	-4.160351	3.36035128	0.9916615

Bonferroni Method

- Let A_1, A_2, \dots, A_n be a set of events
- Bonferroni inequality

$$\Pr(A_1 \cup A_2 \cup \dots \cup A_n) \leq \sum_{i=1}^n \Pr(A_i)$$

- Let A_i be the event that we reject H_{0i} when H_{0i} is true for $i = 1, 2, \dots, n$

$$\Pr(A_i) = \alpha_i$$

Bonferroni Method

- Probability of at least one Type I error

$$\Pr(A_1 \cup A_2 \cup \dots \cup A_n) \leq \sum_{i=1}^n \alpha_i$$

- If $\alpha_i = \alpha^*$ for all i ,

$$\sum_{i=1}^n \alpha_i = n\alpha^*$$

- If we want $\Pr(A_1 \cup \dots \cup A_n) \leq \alpha$, choose $\alpha^* = \alpha/n$
- For ANOVA with K groups, there are $\binom{K}{2}$ tests;
therefore

$$\alpha^* = \frac{\alpha}{\binom{K}{2}}$$

Bonferroni Method: Passive Smoking Example

- $K = 6; \binom{6}{2} = 15$
- $\alpha^* = 0.05/15 = 0.0033$

- Two-sided test,

$$\alpha^*/2 = 0.00167$$

- Rejection region

$$C_\alpha = \{|t_{ij}| > t_{N-K, 1-\alpha^*/2} = t_{1044, 0.9983} = 2.94\}$$

Bonferroni Method

- In SAS proc glm, **means group/bon**;
- In R, `pairwise.t.test(...,p.adj="bonf")`
- Sometimes called the *least-significant difference (LSD)* method (Kleinbaum et al. *Applied Regression Analysis* 3rd edition)
- Applicable well beyond ANOVA
- Choice of $\alpha_i = \alpha / \binom{K}{2}$ for all i is standard, but not necessary

Bonferroni Method

- Definition 12.1: The significance level at which each test or comparison is carried out in an experiment is call the *per comparison error rate (PCER)*
- Bonferroni uses

$$\text{PCER} = \frac{\alpha}{\binom{K}{2}}$$

to ensure

$$\text{PEER} \leq \alpha$$

- Bonferroni-type improvements (Sidak, Holm, Hochberg, Westfall and Young) available; proc glm and proc multtest; beware dependencies in test statistics

Generalizations

- Up to this point we have considered all pairwise comparisons of means
- Other parameter combinations may be of interest
- For instance ...

Factor Level Means

- Single factor level mean

$$\frac{\bar{Y}_{i\cdot} - \mu_i}{\sqrt{\text{MSE}/n_i}} \sim t_{N-K}$$

- $100(1 - \alpha)\%$ CI for μ_i

$$\bar{Y}_{i\cdot} \pm t_{N-K;1-\alpha/2} \sqrt{\text{MSE}/n_i}$$

- Testing $H_0 : \mu_i = c$ vs. $H_A : \mu_i \neq c$

$$t_i = \frac{\bar{Y}_{i\cdot} - c}{\sqrt{\text{MSE}/n_i}} \sim t_{N-K}$$

$$C_\alpha = \{t_i : |t_i| > t_{N-K;1-\alpha/2}\}$$

Linear Combinations and Contrasts

- *Linear combination*

$$L = \sum_{i=1}^K c_i \mu_i$$

- This is a *contrast* if $\sum_i c_i = 0$

- Estimator

$$\hat{L} = \sum_{i=1}^K c_i \bar{Y}_{i\cdot}$$

- Compute CIs and test statistics using

$$\frac{\hat{L} - L}{\sqrt{\text{MSE} \sum_i c_i^2 / n_i}} \sim t_{N-K}$$

Conclusion

- Factor level means, that is, μ_1, μ_2, \dots : Use Bonferroni; in SAS, proc glm/anova with **means group/bon clm**;
- Pairwise comparisons: If balanced, use Tukey; otherwise, if the number of comparisons is not too large and planned *a priori*, use Bonferroni
- Contrasts: Use Scheffé or Bonferroni; for example, multiplicity adjusted CIs for a family of contrasts of the form

$$\hat{L} \pm \sqrt{\text{MSE} \sum_i c_i^2 / n_i} \times \sqrt{(K - 1) F_{K-1, N-K; 1-\alpha}}$$

- Linear combinations: Use Bonferroni

BIOS 662 Fall 2018

Analysis of Variance, Part III

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Diagnostics
- Nonparametric alternative: Kruskal-Wallis

ANOVA: Diagnostics

- Diagnostics discussed in section 10.6 of the text
- Assumptions
 1. Homogeneity of variance
 2. Normality of residual error
 3. Independence of residual error
 4. Linearity

ANOVA: Diagnostics

- Homogeneity of variance
 - Inspect plot of raw data or standard deviations by group means
 - Hartley's and Cochran's test

$$F_{\text{MAX}} = \frac{s_{\max}^2}{s_{\min}^2}, \quad C = \frac{s_{\max}^2}{\sum s_i^2}$$

Tables are given in the Web appendix of the text as “Maximum F Tables” and “Cochran Test Tables”, respectively

- These tests require equal sample size and are sensitive to the normality assumption

ANOVA: Diagnostics

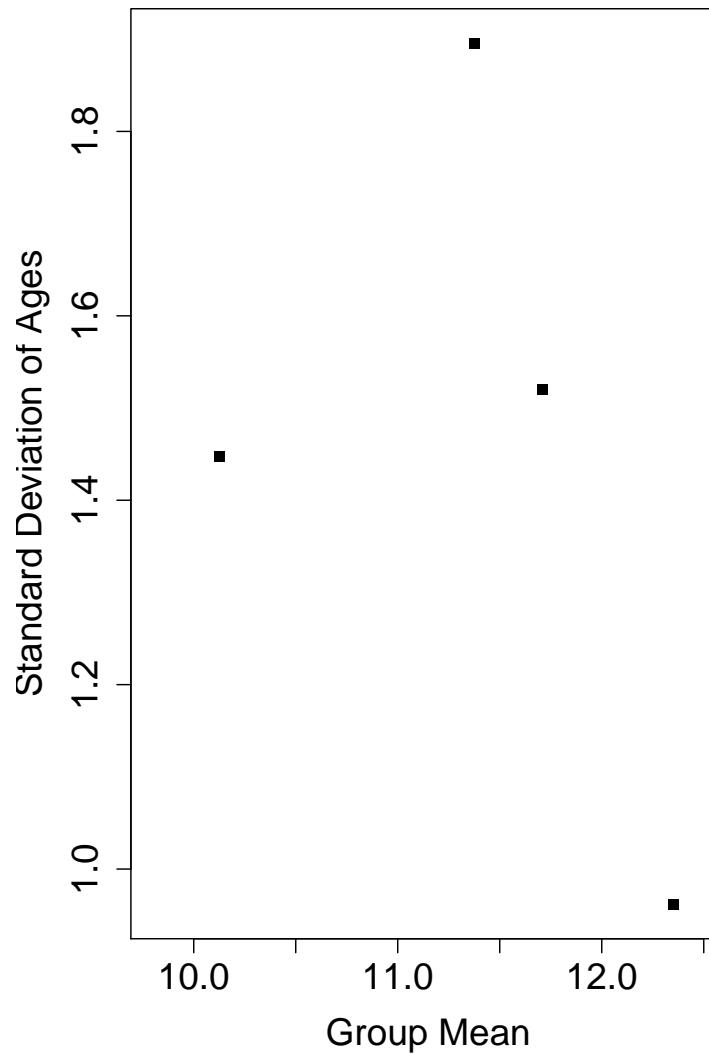
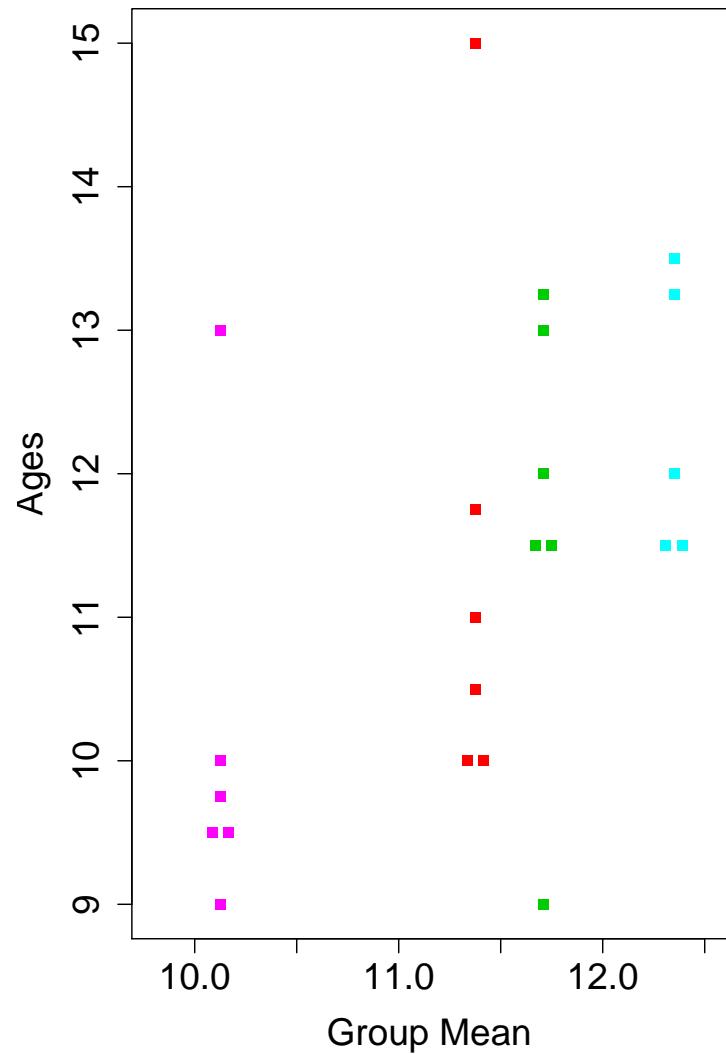
- Homogeneity of variance
 - Modified Levene test (Brown-Forsythe test): apply ANOVA to the absolute deviations from group medians

$$d_{ij} = |Y_{ij} - \tilde{Y}_{i\cdot}|$$

use usual F test; rejection indicates lack of homogeneity
(Ordinary Levene test uses means, not medians)

- Robust to normality; does not require equal sample sizes
- Cf. Chapter 18.2 of Kutner et al. *Applied Linear Statistical Models*, 5th Edition, 2005

Homogeneity of Variance Plot



Modified Levene Test: SAS

```
proc anova; class group; model age=group; means group/hovtest=bf;
```

The ANOVA Procedure

Brown and Forsythe's Test for Homogeneity of age Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
group	3	0.8003	0.2668	0.19	0.9001
Error	19	26.3125	1.3849		

Level of group	N	Mean	Std Dev
active	6	10.1250000	1.44697961
eight	5	12.3500000	0.96176920
no	6	11.7083333	1.52000548
passive	6	11.3750000	1.89571886

Modified Levene Test: R

```
Levene <- function(y, group)
{
  group <- as.factor(group) # precautionary
  medians <- tapply(y, group, median)
  resp <- abs(y - medians[group])
  anova(lm(resp ~ group))[1, 4:5]
}

> Levene(age,group)
      F value Pr(>F)
group 0.1926 0.9001

# Changing anova(lm(resp ~ group))[1, 4:5] to anova(lm(resp ~ group))

> Levene(age,group)
Analysis of Variance Table

Response: resp
          Df  Sum Sq Mean Sq F value Pr(>F)
group       3  0.8003  0.2668  0.1926 0.9001
Residuals 19 26.3125  1.3849
```

ANOVA: Diagnostics for Normality

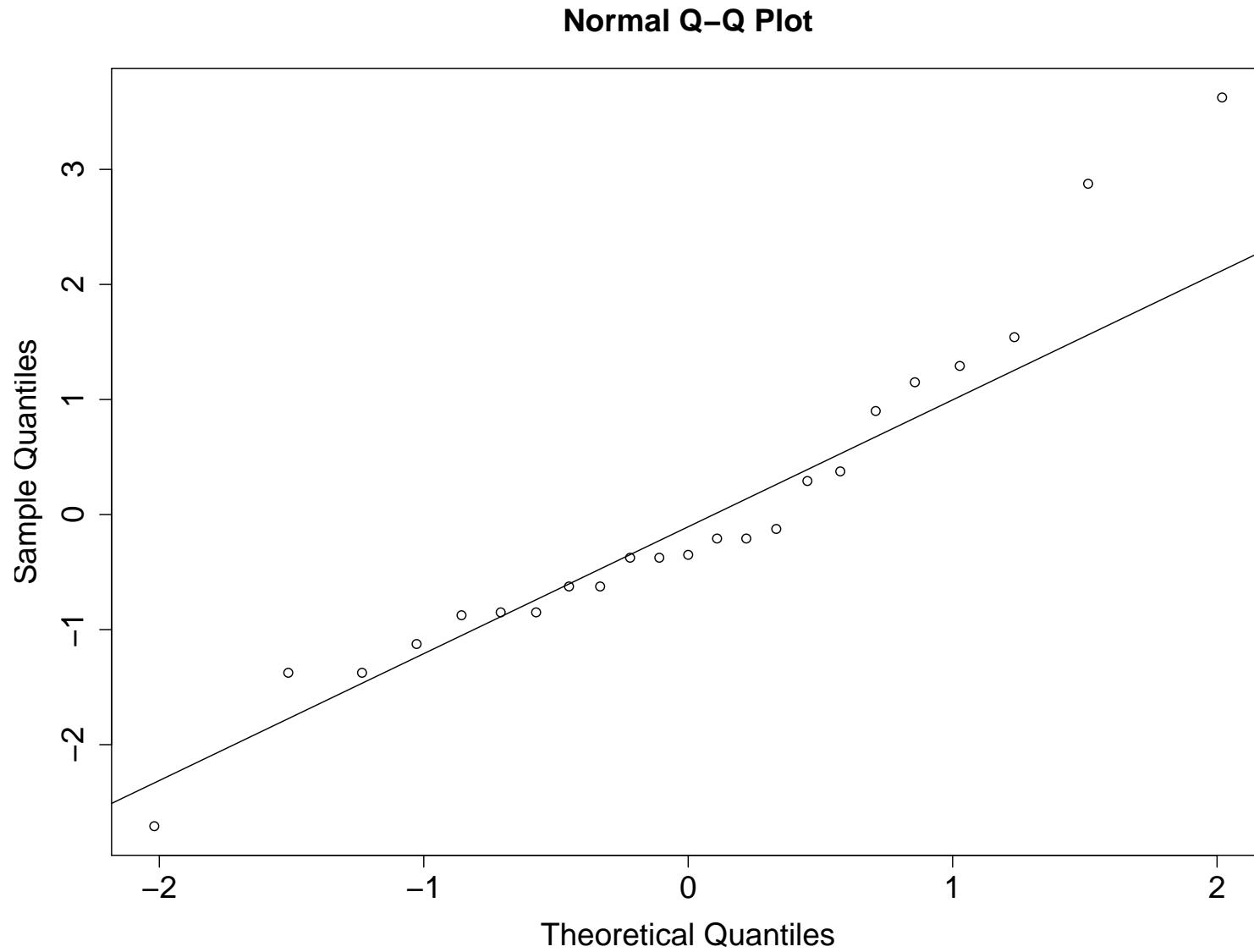
- QQ plot
- K-S GOF test
- Pearson correlation coefficient test:
 - Ordered residuals and expected values under normality
 - Assumption of normality in question if observed correlation is less than or equal to the critical value on the next page

ANOVA: Diagnostics for Normality

- Critical values for $\alpha = 0.05$

N	Crit. val.	N	Crit. val.	N	Crit. val.
5	0.88	10	0.92	24	0.96
6	0.89	12	0.93	30	0.96
7	0.90	15	0.94	40	0.97
8	0.91	20	0.95	50	0.98
9	0.91	22	0.95	100	0.99

ANOVA: Diagnostics for Normality



ANOVA: Diagnostics for Normality in R

```
> group <- as.factor(group)
> av <- aov(age ~ group)
> qq <- qqnorm(av$residuals)
> cor.test(qq$x,qq$y)
```

Pearson's product-moment correlation

```
data: qq$x and qq$y
t = 15.7572, df = 21, p-value = 4.146e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.9070150 0.9832468
sample estimates:
cor
0.9602173
```

ANOVA: Diagnostics

- Remedial measures
 - 1. Normality: appeal to CLT
 - 2. Transformations
 - Plot $(\bar{y}_{i..}, s_i)$, $(\bar{y}_{i..}, s_i^2)$, $(\bar{y}_{i..}^2, s_i)$;
linearity suggests $\log(y)$, \sqrt{y} , $1/y$ transformations, respectively
 - Box-Cox family: minimize SSE (that is, within group SS)
 - 3. Nonparametrics, e.g., Kruskal-Wallis

Box-Cox Transformations

- Family of transformations indexed by λ

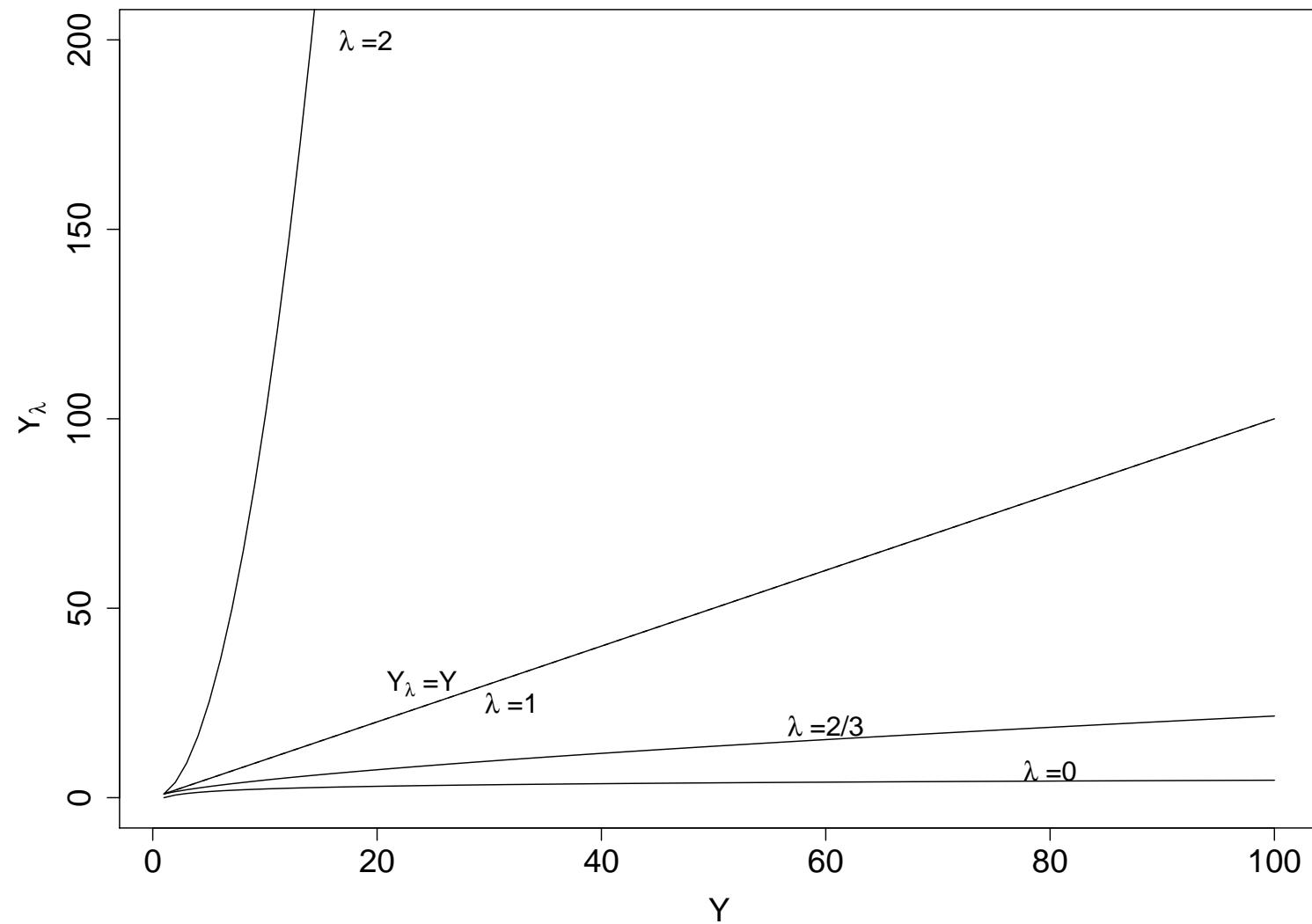
$$Y_\lambda = \begin{cases} k_1(Y^\lambda - 1) & \text{for } \lambda \neq 0 \\ k_2 \log(Y) & \text{for } \lambda = 0 \end{cases}$$

where

$$k_2 = \left(\prod_{i,j} Y_{ij} \right)^{1/N} \quad \text{and} \quad k_1 = \frac{1}{\lambda k_2^{\lambda-1}}$$

- Choose λ that minimizes SSW
- SAS: macro on course website or proc transreg
R: MASS library, function boxcox()

Box-Cox Transformations



Kruskal-Wallis

- Assume

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

for $i = 1, \dots, K; j = 1, \dots, n_i$.

- ϵ_{ij} are independent and identically distributed with mean zero, but not necessarily normal

Kruskal-Wallis

- Same hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \text{ vs. } H_A : \text{at least one inequality}$$

- Pool all N observations and rank from smallest to largest
- Let R_{ij} be the rank of the j^{th} obs in the i^{th} group
- Let $\bar{R}_i = \sum_{j=1}^{n_i} R_{ij}/n_i$ equal the average rank in the i^{th} group
- Let \bar{R} denote the overall average rank. What must this equal?

Kruskal-Wallis

- The Kruskal-Wallis test statistic is

$$T_{\text{KW}} = \frac{12 \sum_{i=1}^K n_i (\bar{R}_i - \bar{R})^2}{N(N+1)}$$

- Equivalently

$$T_{\text{KW}} = \frac{12 \sum_{i=1}^K (\sum_{j=1}^{n_i} R_{ij})^2 / n_i}{N(N+1)} - 3(N+1)$$

- Reject H_0 for large values of T_{KW}

Kruskal-Wallis

- Under H_0 , if the n_i are moderately large (rule of thumb: $n_i \geq 5$), then

$$T_{\text{KW}} \sim \chi^2_{K-1}$$

- If the n_i are small, the exact distribution of T_{KW} can be computed

Kruskal-Wallis: Exact

- There are

$$\binom{N}{n_1 n_2 \cdots n_K} = \frac{N!}{n_1! n_2! n_3! \cdots n_K!}$$

possible ways to assign n_1 ranks to group 1, n_2 ranks to group 2, ...

- Under H_0 each occurs with equal probability
- Suppose $n_1 = 2, n_2 = n_3 = 1$. Then

$$\binom{N}{n_1 n_2 \cdots n_K} = \frac{4!}{2! 1! 1!} = 12$$

Kruskal-Wallis: Exact

R_{1j}	R_{2j}	R_{3j}	$\sum_i R_{i\cdot}^2/n_i$	T_{KW}
1 2	3	4	$9/2+9+16=29.5$	2.7
1 3	2	4	28	1.8
1 4	2	3	25.5	0.3
2 3	1	4	29.5	2.7
2 4	1	3	28	1.8
3 4	1	2	29.5	2.7

k	$\Pr[T_{\text{KW}} = k]$
0.3	$1/6$
1.8	$1/3$
2.7	$1/2$

Kruskal-Wallis with Ties

- If there are ties among the ranks, we use the midrank method as in the Wilcoxon tests
- The KW statistic adjusted for ties is:

$$T_{\text{KWadj}} = \frac{T_{\text{KW}}}{1 - \sum_{i=1}^q (t_i^3 - t_i)/(N^3 - N)}$$

where q is the number of sets of tied observations and t_i is the number of observations in the i^{th} set

- T_{KWadj} will also be approximately χ^2_{K-1}

Kruskal-Wallis: Example

- A study was conducted to compare three doses of aspirin in the treatment of fever in children with the flu
- 15 children with a fever between 100.0 and 100.9 F were randomly assigned to each dose ($n_1 = n_2 = n_3 = 5$; $N = 15$)
- Temperature was measured three hours later
- Let μ_i denote the mean temperature change for dose i
- $H_0 : \mu_1 = \mu_2 = \mu_3$

Kruskal-Wallis: Example cont.

- Distribution of T_{KW} (Owen 1962, page 422; Kruskal, Wallis 1952, JASA, Table 6.1)

k	$\Pr[T_{\text{KW}} \geq k]$
4.50	0.102
4.56	0.100
5.66	0.051
5.78	0.049
7.98	0.010
8.00	0.009

- $C_{0.05} = \{T_{\text{KW}} \geq 5.78\}$

Kruskal-Wallis: Example cont.

Low		Med		High	
ΔTemp	R	ΔTemp	R	ΔTemp	R
2.0	14	0.6	8	1.1	10
1.6	13	1.2	11	-1.0	1
2.1	15	0.5	7	-0.2	3
0.7	9	0.2	4	0.4	6
1.3	12	-0.4	2	0.3	5

- $R_{1\cdot} = 63$, $R_{2\cdot} = 32$, $R_{3\cdot} = 25$

Kruskal-Wallis: Example cont.

- Therefore

$$T_{\text{KW}} = \frac{12(63^2/5 + 32^2/5 + 25^2/5)}{15(16)} - 3(16) = 8.18$$

- Asymptotic p-value

$$\Pr[\chi_2^2 > 8.18] = 0.0167$$

- From Owen table, expect exact p-value < 0.009

Kruskal-Wallis: SAS

```
proc npar1way; class dose; var temp_change; exact wilcoxon;
```

Wilcoxon Scores (Rank Sums) for Variable temp_change

Classified by Variable dose

dose	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
<hr/>					
Low	5	63.0	40.0	8.164966	12.60
Medium	5	32.0	40.0	8.164966	6.40
High	5	25.0	40.0	8.164966	5.00

Kruskal-Wallis Test

Chi-Square	8.1800
DF	2
Asymptotic Pr > Chi-Square	0.0167
Exact Pr >= Chi-Square	0.0081

Kruskal-Wallis: R

```
> kruskal.test(change,dose)
```

```
 Kruskal-Wallis rank sum test
```

```
data: change and dose
```

```
Kruskal-Wallis chi-squared = 8.18, df = 2, p-value = 0.01674
```

Kruskal-Wallis

- Suppose we perform ANOVA with the Y_{ij} replaced by their ranks
- Resulting F test

$$F_R = \frac{(N - K)T_{KW}}{(K - 1)(N - 1 - T_{KW})}$$

- If $K = 2$, the KW test is equivalent to the Wilcoxon rank sum test
- ARE is $3/\pi = 0.955$ compared to the F-test under normality
- For multiple comparisons of means, use Wilcoxon rank sum tests with Bonferroni correction

BIOS 662 Fall 2018

Analysis of Variance, Part IV: A Case Study

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Strategy for Analysis

1. Consider data generation mechanism
2. Analysis plan: Specify model and assumptions;
hypotheses to be tested; diagnostics to be performed
3. Summary statistics, tables, figures
4. Model fitting and diagnostics
5. Inference
6. Sensitivity analysis
7. Conclusions; summary; limitations (e.g., lack of power)

Hypothetical Example

- Survival times of patients following heart transplant surgery based on degree of mismatch (low, medium, high) of the tissue type between donor and recipient

Analysis Plan

- ANOVA model

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

$i = 1, 2, 3$ denoting low, medium, high groups,
respectively

$j = 1, 2, \dots, n_i$ denoting the j^{th} patient in the i^{th} group

$n_1 = 14, n_2 = 13, n_3 = 12$

Y_{ij} survival time in days

μ_i mean survival time in patients with type i mismatch

Analysis Plan cont.

- Primary hypotheses of interest: all pairwise comparisons

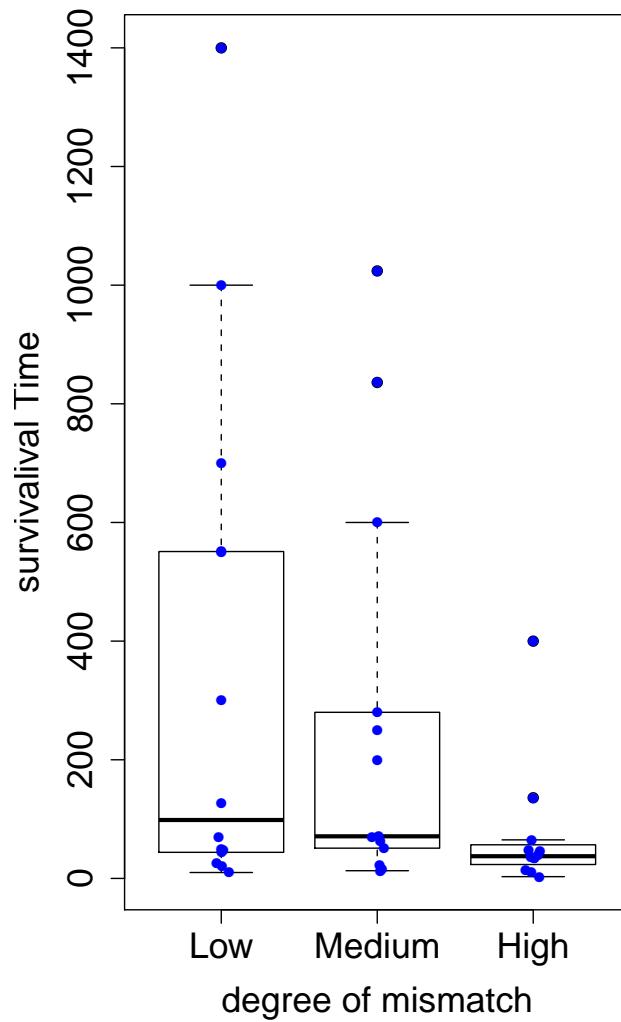
$$H_0 : \mu_1 = \mu_2, \quad H_0 : \mu_1 = \mu_3, \quad H_0 : \mu_2 = \mu_3$$

- Based on power considerations, decide $\alpha = 0.1$
- All pairwise comparisons using Tukey
- Further contrasts, group level means, etc. will be considered hypothesis generating/exploratory; thus no multiplicity adjustment
- Diagnostics, remedial measures, sensitivity analyses

Data

j	Low	Medium	High
1	44	15	3.0
2	551	280	136.0
3	127	1024	65.0
4	1400	836	400.0
5	1000	51	10.4
6	700	600	39.4
7	550	250	33.4
8	300	200	48.4
9	47	22	13.5
10	26	71	34.5
11	50	62	35.5
12	10	69	45.5
13	70	13	
14	20		

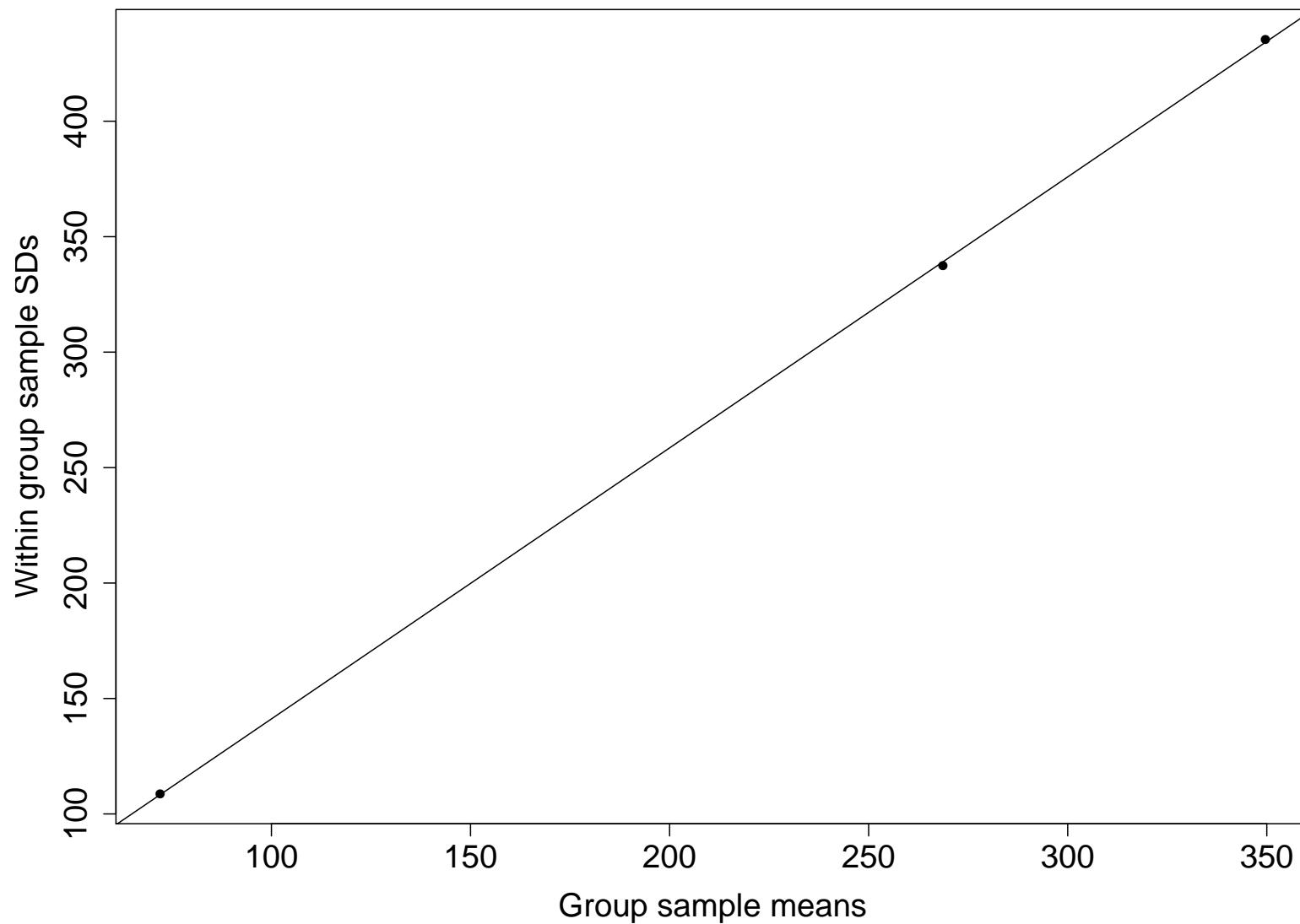
Boxplot With Raw Data



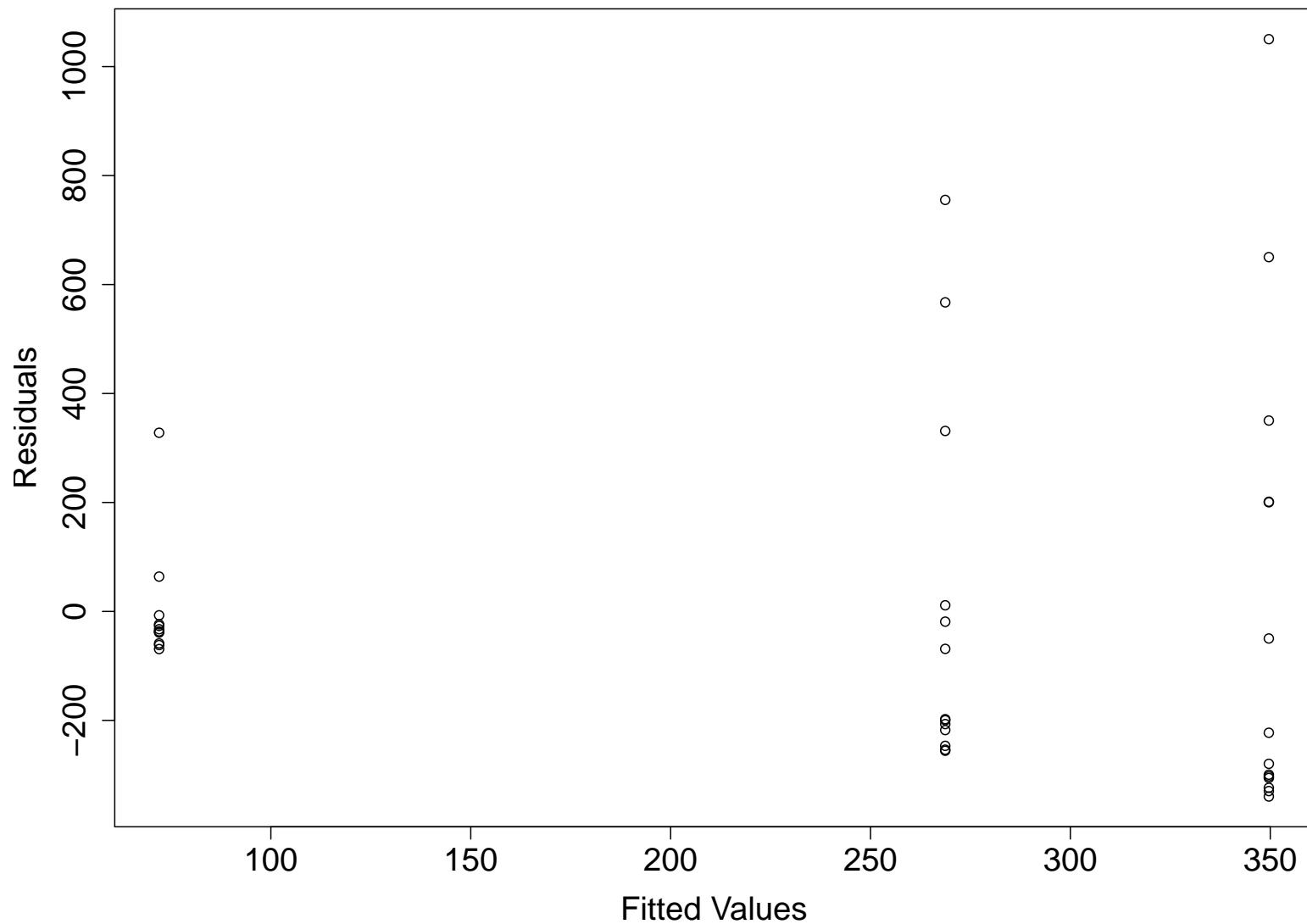
Summary Statistics

	Low	Medium	High
n	14	13	12
Mean	349.6	268.7	72.0
Median	98.5	71.0	37.5
SD	435.31	337.51	108.82
(Min, Max)	(10,1400)	(13, 1042)	(3,400)

Homogeneity of Variance Plot



Residual Plot



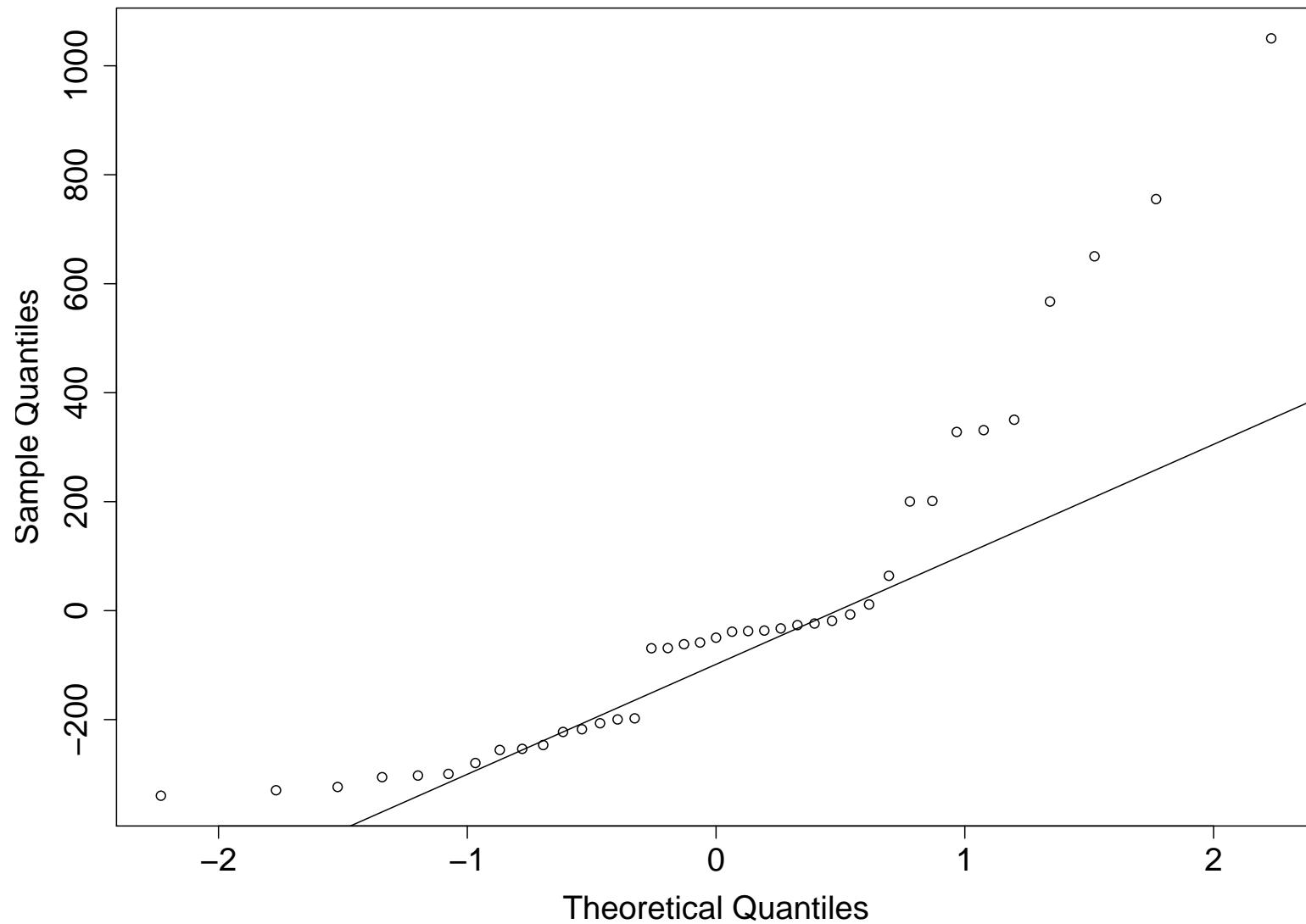
Modified Levene Test

Brown and Forsythe's Test for Homogeneity of survival Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of	Mean	F Value	Pr > F
		Squares	Square		
degree	2	452324	226162	2.46	0.0993
Error	36	3304203	91783.4		

QQ Plot

Normal Q–Q Plot



Box-Cox Transformation

- Recall: Family of transformations indexed by λ

$$Y_\lambda = \begin{cases} k_1(Y^\lambda - 1) & \text{for } \lambda \neq 0 \\ k_2 \log(Y) & \text{for } \lambda = 0 \end{cases}$$

where

$$k_2 = \left(\prod_{i,j} Y_{ij} \right)^{1/N} \quad \text{and} \quad k_1 = \frac{1}{\lambda k_2^{\lambda-1}}$$

- Choose λ that minimizes SSW
- SAS: macro on course website or proc transreg
R: MASS library, function boxcox()

Box-Cox Transformation

```
%macro boxcox(
    resp=,                      /* name of response variable      */
    model=,                     /* independent variables in regression */
    id=,                        /* ID variable for observations   */
    data=_last_,                 /* input dataset                   */
    /* Various other parameters */
    lopower=-2,                  /* low value for power           */
    hipower=2,                   /* high value for power          */
    npower=21,                   /* number of power values in interval */
    conf=0.95);                 /* confidence coefficient of CI on power*/

data case_study;
  infile "anova_case.txt";
  input degree $ survival;

  low=0; medium=0;
  if degree="Low" then low=1;
  if degree="Medium" then medium=1;

*%boxcox(resp=survival,model=low medium,lopower=-2,hipower=2,npower=21);
%boxcox(resp=survival,model=low medium,lopower=-0.7,hipower=0.7,npower=15);
```

Box-Cox Transformation

Box-Cox Power (lambda)	Log Likelihood	Root mean squared error	0.95 Confidence Interval
-0.7	-209.436	214.896	
-0.6	-203.272	183.479	
-0.5	-198.035	160.423	
-0.4	-193.809	143.949	
-0.3	-190.648	132.742	
-0.2	-188.568	125.847	*
-0.1	-187.546	122.593	*
-0.0	-187.530	122.543	<+
0.1	-188.446	125.456	*
0.2	-190.213	131.268	
0.3	-192.746	140.080	
0.4	-195.970	152.149	
0.5	-199.812	167.904	
0.6	-204.212	187.955	
0.7	-209.113	213.124	

Box-Cox Transformation

```
proc transreg data=case_study ss2 pboxcoxtable details;  
  model boxcox(survival) = identity(low medium);
```

Box-Cox Transformation Information for survival

Lambda	R-Square	Log Like
-2.50	0.06	-390.086
-2.00	0.07	-333.778
-1.50	0.07	-280.235
-1.00	0.09	-232.441
-0.50	0.12	-198.035
0.00 +	0.14	-187.530 <
0.50	0.13	-199.812
1.00	0.12	-226.364
1.50	0.10	-261.488
2.00	0.09	-301.910
2.50	0.08	-345.720

< - Best Lambda
* - 95% Confidence Interval
+ - Convenient Lambda

Box-Cox Transformation

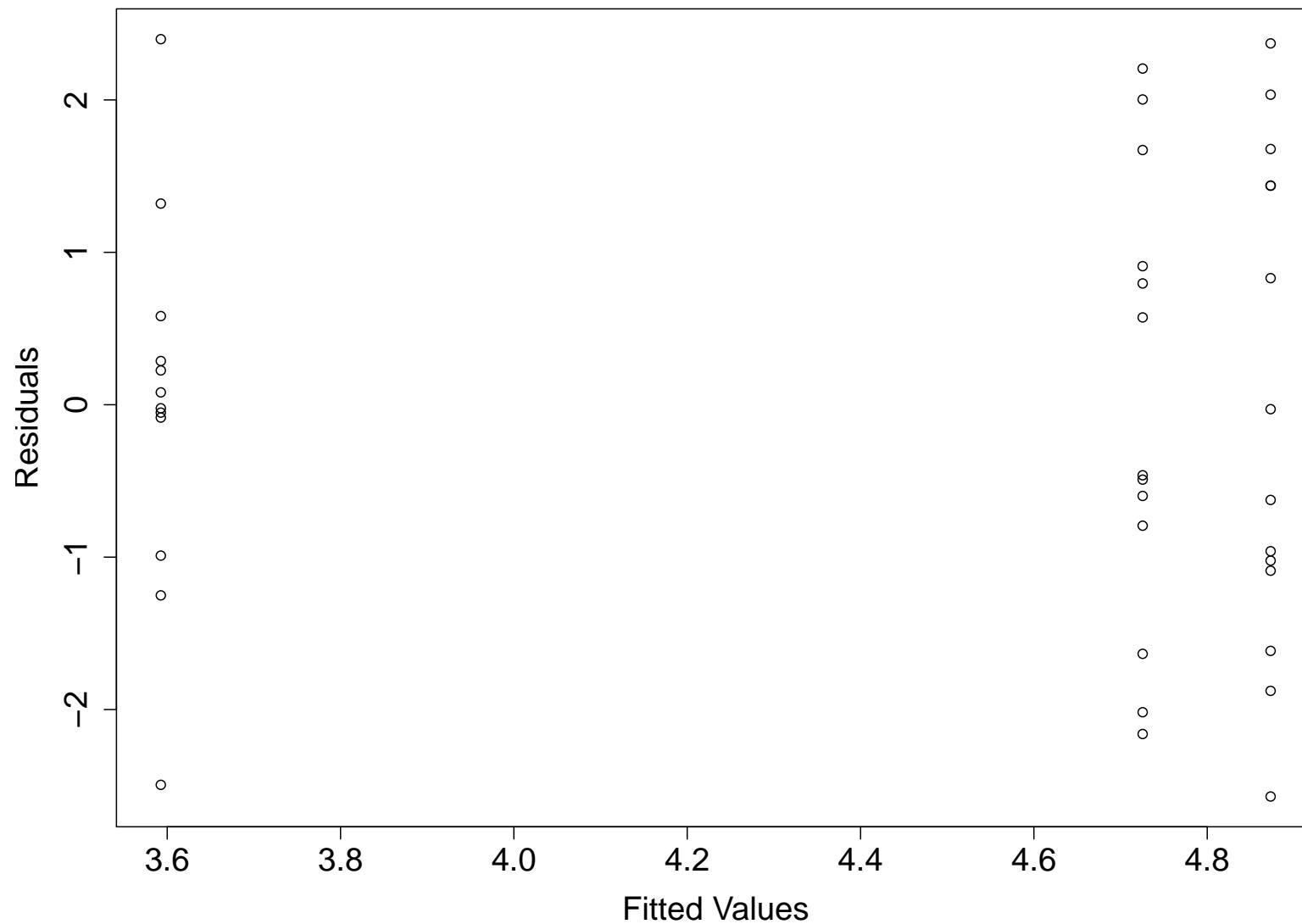
- Choose log transformation
- Run diagnostics again

Modified Levene Test

Brown and Forsythe's Test for Homogeneity of logsurvival Variance
ANOVA of Absolute Deviations from Group Medians

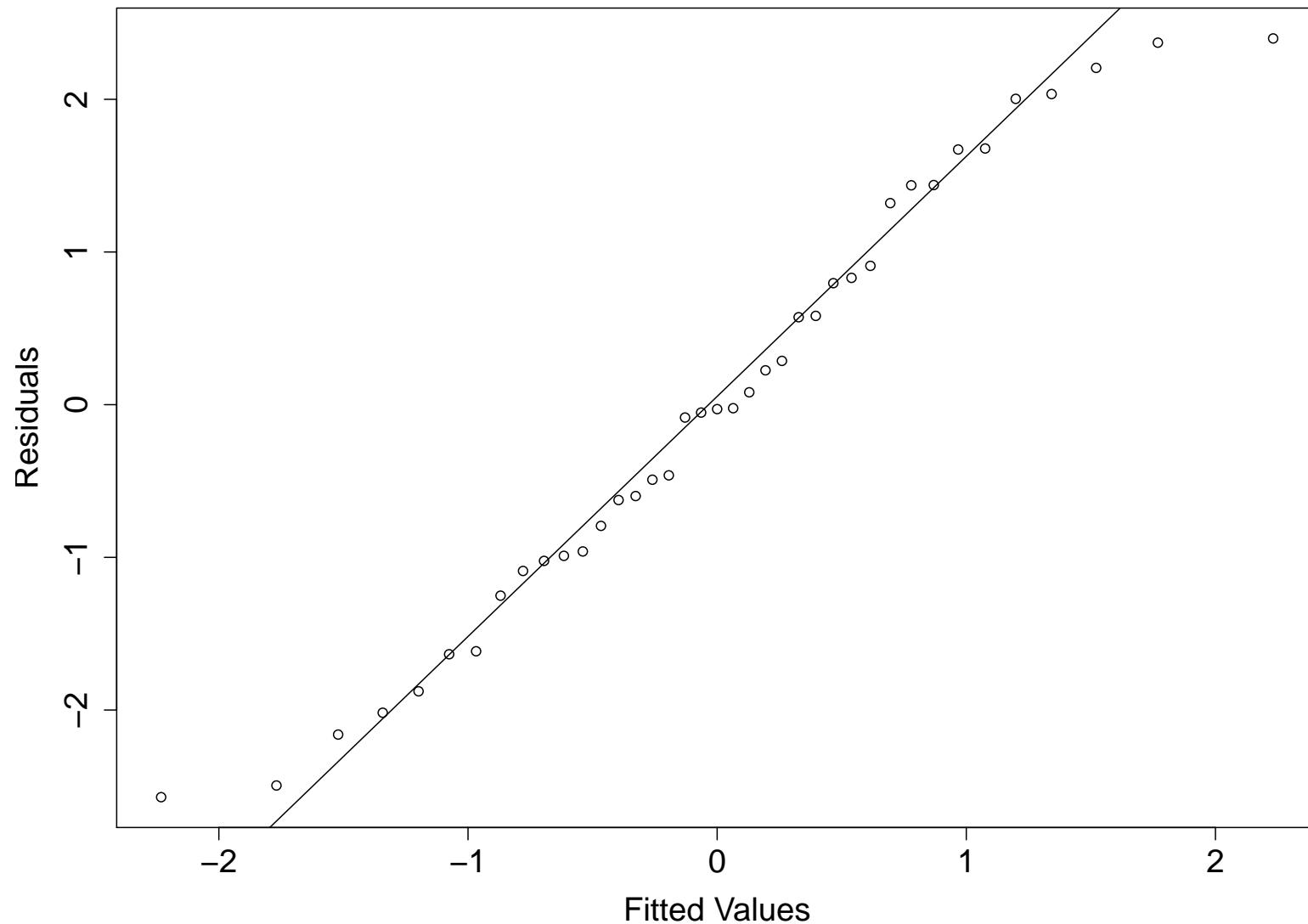
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
degree	2	2.2384	1.1192	1.52	0.2324
Error	36	26.5038	0.7362		

Residual Plot



QQ Plot

Normal Q–Q Plot



Tests for Normality

```
> fit <- aov(log(survival)~degree)
> qq <- qqnorm(fit$residuals,xlab="Fitted Values",ylab="Residuals")
> cor.test(qq$x,qq$y)
```

Pearson's product-moment correlation

sample estimates:

```
cor
0.9882992
```

```
> lillie.test(fit$residuals)
```

Lilliefors (Kolmogorov-Smirnov) normality test

```
data: fit$residuals
D = 0.0799, p-value = 0.766
```

```
> shapiro.test(fit$residuals)
```

```
data: fit$residuals
W = 0.9682, p-value = 0.3294
```

Box-Cox Transformation

- Diagnostics suggest adequate fit after log transformation
- Employ Tukey for all (3) pairwise comparisons of factor level means

Tukey Simultaneous 90% CIs

Tukey's Studentized Range (HSD) Test for logsurvival

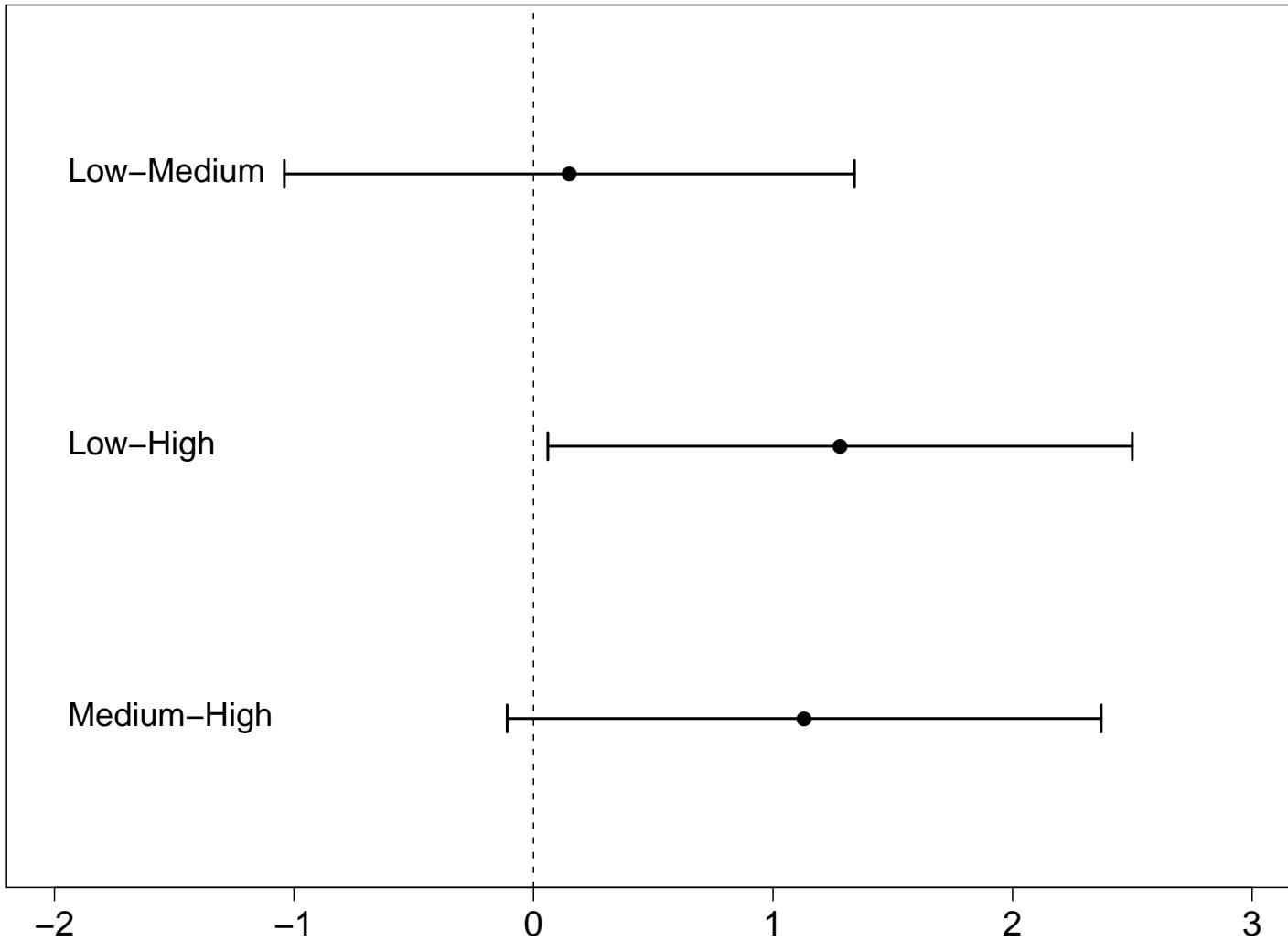
NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.1
Error Degrees of Freedom	36
Error Mean Square	2.132024
Critical Value of Studentized Range	2.99758

Comparisons significant at the 0.1 level are indicated by ***.

Comparison	degree	Difference		
		Between Means	Simultaneous 90% Confidence Limits	
Low - Medium		0.1476	-1.0444	1.3397
Low - High		1.2805	0.0630	2.4980 ***
Medium - Low		-0.1476	-1.3397	1.0444
Medium - High		1.1329	-0.1061	2.3718
High - Low		-1.2805	-2.4980	-0.0630 ***
High - Medium		-1.1329	-2.3718	0.1061

Tukey Simultaneous 90% CIs



Sensitivity Analyses: Bonferroni

Bonferroni (Dunn) t Tests for logsurvival

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Comparisons significant at the 0.1 level are indicated by ***.

Comparison	degree	Difference		
		Between Means	Simultaneous Confidence Limits	90%
Low - Medium		0.1476	-1.0969	1.3922
Low - High		1.2805	0.0094	2.5516 ***
Medium - Low		-0.1476	-1.3922	1.0969
Medium - High		1.1329	-0.1606	2.4263
High - Low		-1.2805	-2.5516	-0.0094 ***
High - Medium		-1.1329	-2.4263	0.1606

Sensitivity Analyses: Wilcoxon Rank Sum

```
wilcox.test(low.survival,med.survival)
```

```
Wilcoxon rank sum test
```

```
W = 93, p-value = 0.943
```

```
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(low.survival,high.survival)
```

```
Wilcoxon rank sum test
```

```
W = 122, p-value = 0.05264
```

```
alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test(med.survival,high.survival)
```

```
Wilcoxon rank sum test
```

```
W = 115, p-value = 0.04571
```

```
alternative hypothesis: true location shift is not equal to 0
```

Contrasts

- It looks like low and medium degrees of mismatch are much better than a high degree of mismatch
- Consider using a contrast to test whether the mean survival time for patients with a high degree of mismatch differs from that of patients in the other two groups
- We want to test

$$H_0 : (\mu_1 + \mu_2)/2 - \mu_3 = 0 \text{ vs. } H_A : (\mu_1 + \mu_2)/2 - \mu_3 \neq 0$$

- Letting $c_1 = 0.5$, $c_2 = 0.5$ and $c_3 = -1$, we have

$$L = \sum_{i=1}^3 c_i \mu_i \quad \text{with} \quad \sum_{i=1}^3 c_i = 0$$

Contrasts cont.

```
data case_study; set case_study;  
logsurvival=log(survival);  
  
if degree='Low' then mismatch=1;  
else if degree='Medium' then mismatch=2;  
else if degree='High' then mismatch=3;  
  
proc glm data=case_study;  
  class mismatch;  
  model logsurvival = mismatch / clparm;  
  estimate 'Low/med vs. high' mismatch 0.5 0.5 -1;  
* or  contrast 'Low/med vs. high' mismatch 0.5 0.5 -1;
```

Standard						
Parameter	Estimate	Error	t Value	Pr > t	95% Conf. Limits	
Low/med vs. high	1.20668	0.50670	2.38	0.0227	0.1791	2.2343
Contrast						
Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F	
Low/med vs. high	1	12.09150	12.09150	5.67	0.0227	

Conclusions

- Conclude there is evidence of a marginally significant difference in mean log survival times between low and high (although not totally consistent across sensitivity analyses)
- How to quantify/interpret on log scale?
- Power of study to detect what size effects?
Alternative analysis to test for trend?

BIOS 662 Fall 2018

Power and Sample Size, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Introduction
- One sample:
 - continuous outcome, Z test
 - continuous outcome, t test
 - binary outcome, Z test
 - binary outcome, exact test

Introduction

- Choosing an appropriate sample size is not just a study design issue, it is an ethical issue
- For a (new) study to be ethical, it must be designed to have sufficient power to detect clinically meaningful differences
- There are ethical issues even if using already-collected data — wasting resources if the sample size is too small
- Power and sample size are mathematically related
- In some situations we can calculate sample sizes explicitly
- In complicated situations, may need to use simulation to determine the sample size

Introduction

- To estimate sample size we need to specify:
 - The study design
 - The significance level α
 - The null hypothesis
 - The test statistic and its distribution
 - The value θ_A that we want to be able to detect
 - The desired power to detect this θ_A
 - More complex models may require specifying other parameters, such as covariances (for measures taken at multiple time-points or if adjusting for confounders)
- These need to be specified when designing the study

Introduction

- How do we decide which values to use?
 - Some are reasonably standard, e.g. α
 - Obtain estimates from pilot studies or studies done elsewhere, e.g. θ_A , variances, covariances
 - θ_A may be what is regarded as the smallest clinically meaningful effect
 - We often calculate sample size for a few representative values of what the underlying parameters might be
 - We often calculate the sample size for two or more choices of the study power — typically 0.8 and 0.9
 - We often choose a few sample sizes and calculate the associated power

Introduction

- These choices need to be made regardless of how the power / sample sizes will be calculated
- These all assume subjects comply with the treatment group to which they are assigned and that we are able to obtain end-point information on all subjects

Introduction

- Read sections 5.8 and 6.3.3 of the text
- In a test of a hypothesis, we are testing whether some population parameter has a particular value

$$H_0 : \theta = \theta_0,$$

where θ_0 is a known constant

- Usually,

$$H_A : \theta \neq \theta_0$$

- Once the data are collected, we compute a statistic related to θ , say $S(\hat{\theta})$
- $S(\hat{\theta})$ is a random variable, because it is computed from a sample (and hence it has a probability distribution)

Power

$$\Pr[\text{Type I error}] = \alpha = \Pr[S(\hat{\theta}) \in C_\alpha \mid H_0]$$

$$\Pr[\text{Type II error}] = \beta = \Pr[S(\hat{\theta}) \notin C_\alpha \mid H_A]$$

$$\text{Power} = 1 - \beta = \Pr[S(\hat{\theta}) \in C_\alpha \mid H_A]$$

One Sample Z Test

- Example: One sample test
- Study: Collect data for a continuous outcome Y on N individuals
- $E(Y) = \mu$, $\text{Var}(Y) = \sigma^2$

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_A : \mu > \mu_0$$

$$S(\hat{\theta}) = Z = \frac{\bar{Y} - \mu_0}{\sigma / \sqrt{N}}$$

One Sample Z Test

$$\Pr[S(\hat{\theta}) \in C_\alpha \mid H_0] = \Pr[Z > z_{1-\alpha} \mid H_0] = \alpha$$

$$\Pr[S(\hat{\theta}) \in C_\alpha \mid H_A] = \Pr[Z > z_{1-\alpha} \mid H_A] = 1 - \beta$$

- Choose a value $\mu_A \in H_A$
- The question of interest is: What sample size do we need in order to have power $1 - \beta$ to detect this alternative?
- The sample size depends on the particular choice of μ_A

One Sample Z Test

- Under $H_A : \mu = \mu_A$

$$Z' = \frac{\bar{Y} - \mu_A}{\sigma/\sqrt{N}} \sim N(0, 1)$$

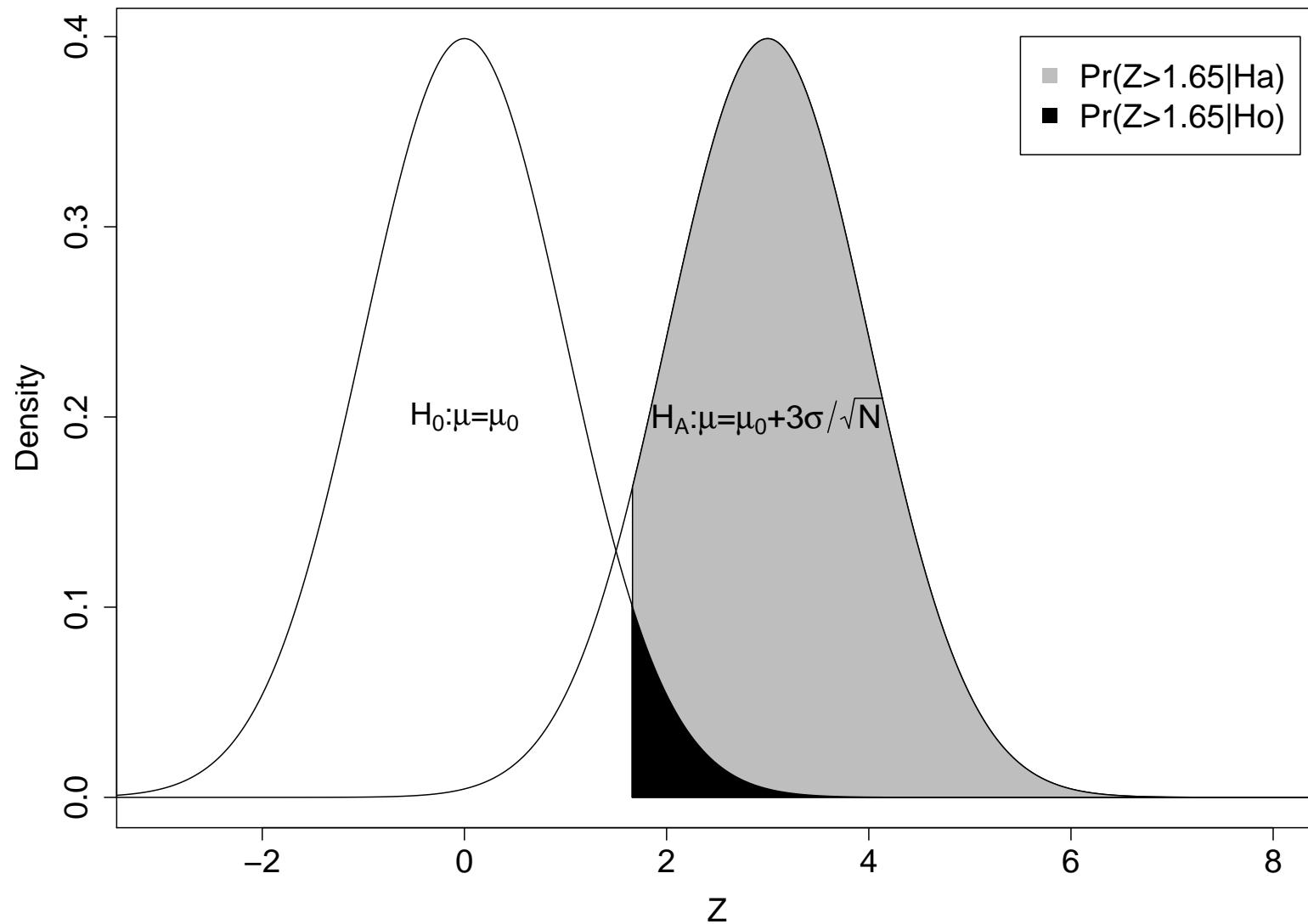
- Note

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{N}} = Z' + \frac{\mu_A - \mu_0}{\sigma/\sqrt{N}}$$

- Thus

$$Z \sim N\left(\frac{\mu_A - \mu_0}{\sigma/\sqrt{N}}, 1\right)$$

Distribution of Z -statistic under H_0 and H_A



One Sample Z Test

- Power is given by

$$\begin{aligned}1 - \beta &= \Pr[Z > z_{1-\alpha} \mid \mu = \mu_A] \\&= \Pr[Z' > z_{1-\alpha} + \frac{\sqrt{N}(\mu_0 - \mu_A)}{\sigma} \mid \mu = \mu_A]\end{aligned}$$

So

$$\beta = \Pr[Z' \leq z_{1-\alpha} + \frac{\sqrt{N}(\mu_0 - \mu_A)}{\sigma} \mid \mu = \mu_A]$$

- Therefore

$$z_{1-\alpha} + \frac{\sqrt{N}(\mu_0 - \mu_A)}{\sigma} = z_\beta = -z_{1-\beta}$$

One Sample Z Test

- Equivalently

$$N = \frac{\sigma^2(z_{1-\alpha} + z_{1-\beta})^2}{(\mu_0 - \mu_A)^2}$$

- For a two sided test

$$N = \frac{\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_0 - \mu_A)^2}$$

One Sample Z Test

- Equivalent form

$$N = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta} \right)^2$$

where

$$\Delta = \frac{|\mu_0 - \mu_A|}{\sigma}$$

- Δ is called the *standardized distance* or *difference*

One Sample Z Test

- To apply this formula, we need to know α , β , $|\mu_0 - \mu_A|$, and σ^2
- Example: A study is planned to determine the effect of a drug on blood pressure.
- BP will be measured, a drug administered, and BP measured again 2 hours later
- $Y = (\text{BP}_{\text{after}} - \text{BP}_{\text{before}})$

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_A : \mu \neq 0$$

One Sample Z Test

- Choose $\alpha = 0.05$, $1 - \beta = 0.9$
- Estimate of σ^2 : Need data from the literature or a pilot study; note that we need the variance of the difference (after – before)
- The choice of μ_A is a subject-matter decision. In this example, a drug that changes BP by just 1 mmHg would not be of practical importance, but a drug that changes BP by 5 mmHg might be of interest

One Sample Z Test

- Suppose $\alpha = 0.05$, $1 - \beta = 0.9$, $\sigma^2 = 225$, and $\mu_A = 5$

$$N = \frac{225(1.96 + 1.28)^2}{5^2} = 94.5 \approx 95$$

- Often compute N for various different values of α , β , σ^2 , and μ_A
- Note that $(1.96 + 1.28)^2 \approx 10.5$, so that

$$N \approx \frac{10.5}{\Delta^2}$$

One Sample Z Test

α	$1 - \beta$	σ^2	μ_A	N	α	$1 - \beta$	σ^2	μ_A	N
0.05	0.90	225	5	95	0.01	0.90	225	5	133
		225	6	66			225	6	93
		256	5	108			256	5	151
		256	6	75			256	6	105
0.05	0.80	225	5	71	0.01	0.80	225	5	105
		225	6	49			225	6	73
		256	5	81			256	5	119
		256	6	56			256	6	83

One Sample Z Test

$$\alpha \downarrow \Rightarrow N \uparrow$$

$$1 - \beta \uparrow \Rightarrow N \uparrow$$

$$\sigma^2 \uparrow \Rightarrow N \uparrow$$

$$|\mu_0 - \mu_A| \downarrow \Rightarrow N \uparrow$$

One Sample Z Test

- Sometimes N is fixed and we want to estimate the power of the test

$$N = \frac{\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_0 - \mu_A)^2}$$

$$z_{1-\beta} = \frac{|\mu_0 - \mu_A| \sqrt{N}}{\sigma} - z_{1-\alpha/2}$$

$$1 - \beta = \Phi\left(\frac{|\mu_0 - \mu_A| \sqrt{N}}{\sigma} - z_{1-\alpha/2}\right)$$

One Sample Z Test: Example

- An investigator says that the budget can accommodate just 50 patients
- Suppose $\alpha = 0.05$, $\sigma^2 = 225$, $|\mu_0 - \mu_A| = 5$
- Then

$$z_{1-\beta} = \frac{5\sqrt{50}}{15} - 1.96 = 0.40$$

$$1 - \beta = 0.65$$

One Sample t Test

- In practice, σ is not known, so we use a t test instead of a Z test
- The previous results should be viewed as approximations in this case
- Now derive the power for a t test
- Need the following result: If $U \sim N(\lambda, 1)$ and $V \sim \chi^2_\nu$ with $U \perp V$, then

$$\frac{U}{\sqrt{V/\nu}} \sim t_{\nu, \lambda}$$

that is, a non-central t distribution with ν degrees of freedom and non-centrality parameter λ

One Sample t Test

- Consider $H_0 : \mu = 0$ versus $H_A : \mu = \mu_A$ for $\mu_A \neq 0$

- Under H_A ,

$$\bar{Y} \sim N(\mu_A, \sigma^2/N)$$

- Thus

$$\frac{\bar{Y}\sqrt{N}}{\sigma} \sim N\left(\mu_A \frac{\sqrt{N}}{\sigma}, 1\right)$$

- Recall

$$\frac{(N-1)s^2}{\sigma^2} \sim \chi_{N-1}^2$$

One Sample t Test

- Because $\bar{Y} \perp s^2$,

$$T = \frac{\bar{Y}}{s/\sqrt{N}} \sim t_{N-1, \lambda}$$

where $\lambda = \mu_A \sqrt{N}/\sigma$

- So the power of a two-sided t test for $\mu_A > 0$ is

$$\Pr[T \geq t_{N-1; 1-\alpha/2}]$$

where $T \sim t_{N-1, \mu_A \sqrt{N}/\sigma}$

One Sample t Test: R

```
# by hand
```

```
> 1-pt(qt(0.975,49), 49, 5*sqrt(50)/15)
[1] 0.6370846
```

```
> power.t.test(n=50, sd=15, delta=5, type="one.sample")
```

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

```
n = 50
delta = 5
sd = 15
sig.level = 0.05
power = 0.6370846
alternative = two.sided
```

One Sample t Test: SAS

```
proc power;  
    onesamplemeans  
    mean      = 5  
    ntotal    = 50  
    stddev   = 15  
    power    = .;
```

One-sample t Test for Mean

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Mean	5
Standard Deviation	15
Total Sample Size	50
Number of Sides	2
Null Mean	0
Alpha	0.05

Computed Power

0.637

One Sample Z Test: Binary Outcome

- Null hypothesis

$$H_0 : \pi = \pi_0$$

- Test statistic

$$Z = \frac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/N}}$$

- Sample size formula

$$N = \frac{\left(z_{1-\alpha/2} \sqrt{\pi_0(1 - \pi_0)} + z_{1-\beta} \sqrt{\pi_A(1 - \pi_A)} \right)^2}{(\pi_A - \pi_0)^2}$$

One Sample Z Test, Binary Outcome: Example

- Study of risk of breast cancer in siblings.
- Prevalence in population of women 50-54 years old: 2%
- Plan to sample sisters of women with breast cancer and test $H_0 : \pi = 0.02$ versus $H_A : \pi \neq 0.02$
- Suppose $\alpha = 0.05$, $1 - \beta = 0.9$, $\pi_A = 0.05$
- Then

$$N = \frac{\left(1.96\sqrt{0.02(0.98)} + 1.28\sqrt{0.05(0.95)}\right)^2}{(0.05 - 0.02)^2} = 340.24$$

- Round up to 341

One Sample Z Test With Binary Outcome: SAS

```
proc power;  
    onestepfreq test = z  
    method = normal  
    nullp  = 0.02  
    p      = 0.05  
    power  = 0.9  
    ntotal = .;
```

The POWER Procedure
Z Test for Binomial Proportion

Fixed Scenario Elements

Method	Normal approximation
Null Proportion	0.02
Binomial Proportion	0.05
Nominal Power	0.9
Number of Sides	2
Alpha	0.05

Actual	N
Power	Total
0.900	341

One Sample Exact Test: Binary Outcome

- What is the power of the exact test?
- Let Y , the number of successes, be the test statistic.
- Under H_0 , $Y \sim \text{Binomial}(N, \pi_0)$
Under H_A , $Y \sim \text{Binomial}(N, \pi_A)$
- Power

$$\Pr[Y \geq y_{1-\alpha/2} \mid \pi = \pi_A] + \Pr[Y \leq y_{\alpha/2} \mid \pi = \pi_A]$$

where $y_{\alpha/2}$ and $y_{1-\alpha/2}$ are determined as in the notes
on “Count Data”

Exact Test, Binary Outcome: Example

- Suppose $N = 20$, $\pi_0 = 0.2$, $\pi_A = 0.5$, $\alpha = 0.05$
- What is exact power of a two-sided exact test?
- Based on the CDF of $\text{Binomial}(20, 0.2)$, choose
 $y_{\alpha/2} = 0$ and $y_{1-\alpha/2} = 9$
- Then the power is

$$\Pr[Y \geq 9 \mid \pi = 0.5] + \Pr[Y \leq 0 \mid \pi = 0.5] = 0.748$$

Exact Test With Binary Outcome: SAS

```
proc power;  
    onesamplefreq test = exact  
    method = exact  
    nullp  = 0.2  
    p      = 0.5  
    power  = .  
    ntotal = 20;
```

Computed Power

Lower	Upper		
Crit	Crit	Actual	
Val	Val	Alpha	Power
0	9	0.0215	0.748

BIOS 662 Fall 2018

Power and Sample Size, Part II

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Two sample: Continuous
- Two sample: Binary
- Case-control studies
- Estimating power with simulations

Two Sample Test: Continuous Outcome

- Hypotheses

$$H_0 : \mu_1 = \mu_2 \text{ versus } H_A : \mu_1 \neq \mu_2$$

- Assume homogeneity of variance, σ^2 known,
normality/CLT
- Then

$$N = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\mu_1 - \mu_2)^2} = 2 \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta} \right)^2$$

- Note that there are N observations in *each* group so
that the total sample size is $2N$

Two Sample Test: Example

- A drug company wants to compare 2 drugs for lowering LDL-cholesterol
- Previous studies have found $\sigma^2 = 25^2 = 625$
- A difference of 15 mg/dl is considered to be clinically meaningful
- For $\alpha = 0.05$ (2-sided) and $1 - \beta = 0.9$
$$N = \frac{2(625)(1.96 + 1.28)^2}{225} \approx 59$$
- So 118 patients are needed for the study

Two Sample Test: σ Unknown

- What if the variance is not known?
- For $N_1 = N_2 = N$, one can show that

$$\frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{\frac{2}{N}}} \sim t_{2N-2, \lambda}$$

where

$$\lambda = \Delta \sqrt{N/2}$$

Two Sample Test, σ Unknown: R

```
# by hand  
> 1-pt(qt(0.975,116), 116, 15/25*sqrt(59/2))  
[1] 0.8982732
```

```
> power.t.test(59, delta=15, sd=25)
```

Two-sample t test power calculation

```
n = 59  
delta = 15  
sd = 25  
sig.level = 0.05  
power = 0.8982732  
alternative = two.sided
```

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

Two Sample Test, σ Unknown: SAS

```
proc power;  
  twosamplemeans  
  meandiff = 15  
  ntotal    = 118  
  stddev    = 25  
  power     = .;
```

Two-sample t Test for Mean Difference

Distribution	Normal
Method	Exact
Mean Difference	15
Standard Deviation	25
Total Sample Size	118
Number of Sides	2
Null Difference	0
Alpha	0.05

Computed Power

0.898

Two Sample Test, σ Unknown

- Given β , solve for N

$$1 - \beta = \Pr[T \geq t_{2N-2,0;1-\alpha/2}]$$

where $T \sim t_{2N-2,\Delta\sqrt{N/2}}$

- For example, suppose $\beta = 0.1$, $\Delta = 0.5$; numerical search in R:

```
> N <- 50; 1-pt(qt(0.975,2*N-2), 2*N-2, 1/2*sqrt(N/2))
[1] 0.6968888
> N <- 90; 1-pt(qt(0.975,2*N-2), 2*N-2, 1/2*sqrt(N/2))
[1] 0.9155872
> N <- 86; 1-pt(qt(0.975,2*N-2), 2*N-2, 1/2*sqrt(N/2))
[1] 0.9032299
> N <- 85; 1-pt(qt(0.975,2*N-2), 2*N-2, 1/2*sqrt(N/2))
[1] 0.899894
```

- So $N = 86$

Two Sample Test, σ Unknown: R

```
> power.t.test(power=0.9, delta=0.5)
```

```
Two-sample t test power calculation
```

```
n = 85.03129
delta = 0.5
sd = 1
sig.level = 0.05
power = 0.9
alternative = two.sided
```

Two Sample Test, σ Unknown: SAS

```
proc power;  
  twosamplemeans  
  meandiff = 15  
  ntotal    = .  
  stddev    = 30  
  power     = 0.9;
```

Two-sample t Test for Mean Difference

Distribution	Normal
Method	Exact
Mean Difference	15
Standard Deviation	30
Nominal Power	0.9
Number of Sides	2
Null Difference	0

Computed N Total

Actual	N
Power	Total
0.903	172

Two Sample Test: Binary Outcome

- Hypotheses

$$H_0 : \pi_1 = \pi_2 \text{ versus } H_A : \pi_1 \neq \pi_2$$

- Then

$$N \approx \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\pi_1 - \pi_2)^2}$$

where

$$\sigma^2 = (\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2))/2$$

see page 161 of the text

- Again there are N observations in each group so that the total sample size is $2N$

Two Sample Test: Binary Outcome

- Suppose $\pi_1 = 0.2727$, $\pi_2 = 0.2$, $\alpha = 0.05$ (two-sided), $1 - \beta = 0.9$. Then

$$N \approx \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{(\pi_1 - \pi_2)^2} = 712$$

- SAS

```
proc power;
  twosamplefreq
    refp   = 0.2
    pdiff  = 0.0727
    ntotal = .
    power  = 0.9;
```

Two Sample Test: Binary Outcome

Pearson Chi-square Test for Two Proportions

Fixed Scenario Elements

Distribution	Asymptotic normal
Method	Normal approximation
Reference (Group 1) Proportion	0.2
Proportion Difference	0.0727
Nominal Power	0.9
Number of Sides	2
Null Proportion Difference	0
Alpha	0.05
Group 1 Weight	1
Group 2 Weight	1

Computed N Total

Actual	N
Power	Total
0.900	1432

Two Sample Test: Binary Outcome

- Why the difference? SAS uses a different approximation, which we now derive (cf. Fleiss, 1981)
- For $N_1 = N_2 = N$, Pearson's chi-square test statistic is equivalent to

$$Z = \frac{p_2 - p_1}{\sqrt{2\bar{p}\bar{q}/N}}$$

where $\bar{p} = (p_1 + p_2)/2$, $\bar{q} = 1 - \bar{p}$

- Without loss of generality, consider the alternative $\pi_2 - \pi_1 = \delta_A > 0$.

Two Sample Test: Binary Outcome

- Power to detect δ_A using a two-sided test is

$$\begin{aligned}\Pr[Z > z_{1-\alpha/2} \mid \delta_A] + \Pr[Z < z_{\alpha/2} \mid \delta_A] \\ \approx \Pr[Z > z_{1-\alpha/2} \mid \delta_A]\end{aligned}$$

- We need to know the distribution of Z under H_A

$$E(p_2 - p_1) = \delta_A$$

$$\text{Var}(p_2 - p_1) = \frac{\pi_2(1 - \pi_2)}{N} + \frac{\pi_1(1 - \pi_1)}{N}$$

Two Sample Test: Binary Outcome

- So

$$\begin{aligned}1 - \beta &= \Pr \left[\frac{p_2 - p_1}{\sqrt{\frac{2\bar{p}\bar{q}}{N}}} > z_{1-\alpha/2} \mid \delta_A \right] \\&= \Pr \left[p_2 - p_1 > z_{1-\alpha/2} \sqrt{\frac{2\bar{p}\bar{q}}{N}} \mid \delta_A \right] \\&= \Pr \left[\frac{(p_2 - p_1) - \delta_A}{\sqrt{\text{Var}(p_2 - p_1)}} > \frac{z_{1-\alpha/2} \sqrt{\frac{2\bar{p}\bar{q}}{N}} - \delta_A}{\sqrt{\text{Var}(p_2 - p_1)}} \mid \delta_A \right]\end{aligned}$$

Two Sample Test: Binary Outcome

- Implying

$$-z_{1-\beta} = \frac{z_{1-\alpha/2} \sqrt{2\bar{p}\bar{q}/N} - \delta_A}{\sqrt{\text{Var}(p_2 - p_1)}}$$

- Using $\bar{p}\bar{q} \approx \bar{\pi}(1 - \bar{\pi})$ where $\bar{\pi} = (\pi_1 + \pi_2)/2$ yields

$$z_{1-\beta} \sqrt{\text{Var}(p_2 - p_1)} + z_{1-\alpha/2} \sqrt{2\bar{\pi}(1 - \bar{\pi})/N} = \delta_A$$

Two Sample Test: Binary Outcome

- Therefore

$$\frac{z_{1-\beta} \sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)} + z_{1-\alpha/2} \sqrt{2\bar{\pi}(1-\bar{\pi})}}{\delta_A} = \sqrt{N}$$

- Thus, the sample size required per arm to detect δ_A with power $1 - \beta$ is

$$\frac{(z_{1-\beta} \sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)} + z_{1-\alpha/2} \sqrt{2\bar{\pi}(1-\bar{\pi})})^2}{\delta_A^2}$$

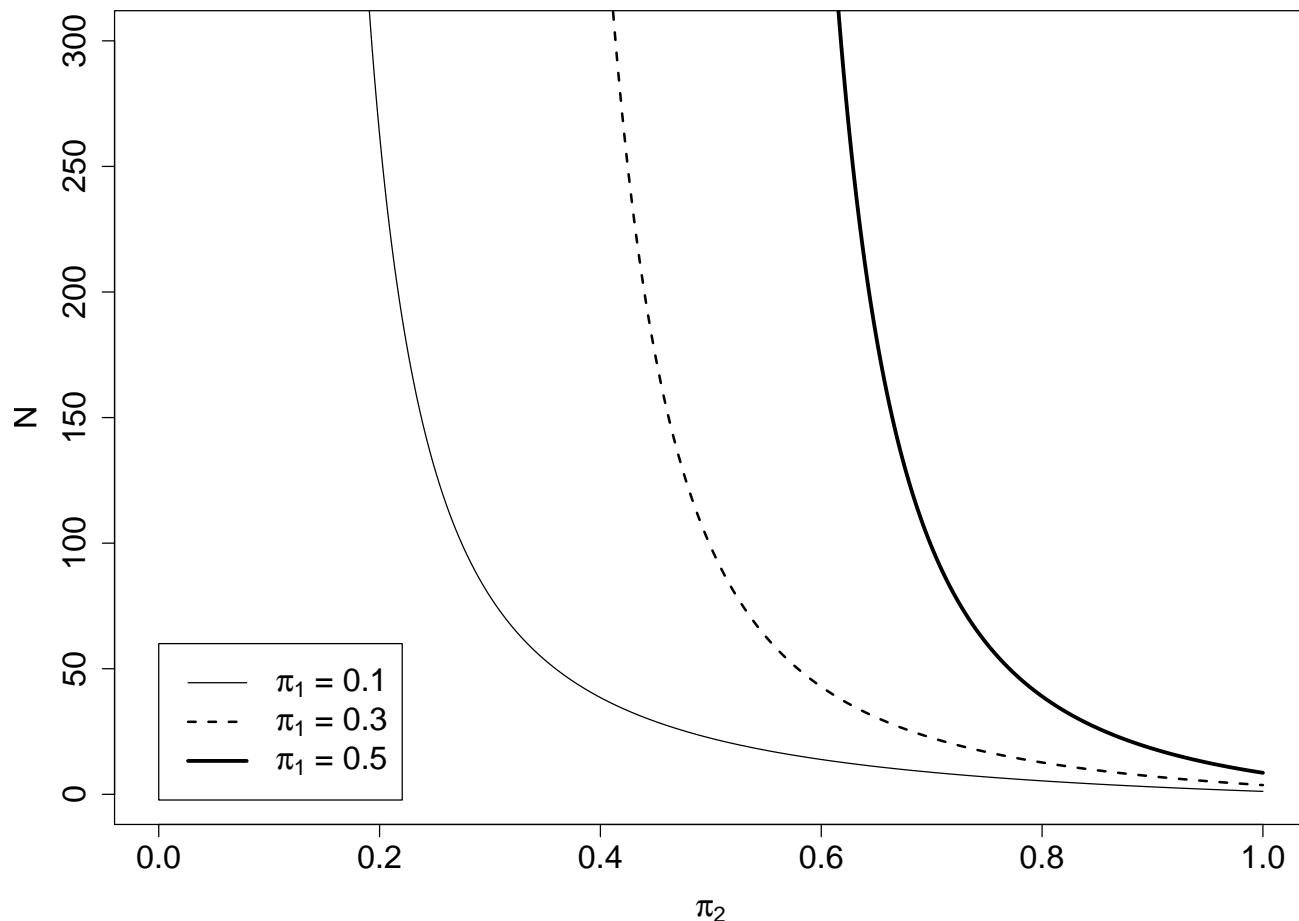
Two Sample Test: Binary Outcome

- In R by hand

```
# sample size formula for comparing two  
# binomial proportions based on Fleiss (second edition) page 41  
  
ss_fleiss <- function(pi1,pi2,alpha,power){  
  q1 <- 1-pi1  
  q2 <- 1-pi2  
  pbar <- (pi1+pi2)/2  
  qbar <- 1-pbar  
  num <- qnorm(1-alpha/2)*sqrt(2*pbar*qbar)+qnorm(power)*sqrt(pi1*q1+pi2*q2)  
  den <- (pi2-pi1)  
  (num/den)^2  
}  
  
> ss_fleiss(0.2,0.2727,0.05,0.9)  
[1] 715.5618
```

Graphical Summary

Sample size (per arm) for comparing π_1 against π_2 with $\alpha = 0.05$ (one-sided) and 90% power



Case-Control: Binary Exposure

- Hypotheses

$$H_0 : \text{OR} = 1 \quad \text{vs.} \quad H_A : \text{OR} \neq 1$$

$$\text{OR} = \frac{\text{odds}(\text{disease} \mid \text{exposed})}{\text{odds}(\text{disease} \mid \text{unexposed})}$$

- Recall

		Disease	No disease
Exposed	Disease	π_{11}	π_{12}
	No disease	π_{21}	π_{22}

Case-Control: Binary Exposure

$$\begin{aligned} \text{OR} &= \frac{\text{odds(disease} \mid \text{exposed})}{\text{odds(disease} \mid \text{unexposed})} \\ &= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \\ &= \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} \\ &= \frac{\text{odds(exposed} \mid \text{disease})}{\text{odds(exposed} \mid \text{no disease})} \end{aligned}$$

Case-Control: Binary Exposure

- Hypotheses

$$H_0 : \text{OR} = 1 \quad \text{vs.} \quad H_A : \text{OR} \neq 1$$

- From the previous page

$$\text{OR} = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)}$$

where $\pi_1 = \Pr(\text{exposed} \mid \text{case})$,

$\pi_2 = \Pr(\text{exposed} \mid \text{control})$

- For specified OR and π_2 we can determine π_1

$$\pi_1 = \frac{\pi_2 \text{OR}}{1 + \pi_2(\text{OR} - 1)}$$

Case-Control, Binary Exposure: Example

- Cases: Neural tube defect babies
Controls: Normal babies
Exposure: Self reported dieting to lose weight during first trimester
- It is estimated that 20% of women will diet to lose weight during pregnancy; $\pi_2 = 0.2$
- The investigator wants to be able to detect $OR = 1.5$

Case-Control, Binary Exposure: Example cont.

- First determine the corresponding value of π_1 :

$$\pi_1 = \frac{0.2(1.5)}{1 + 0.2(0.5)} = 0.2727$$

- For $\pi_1 = 0.2727$, $\pi_2 = 0.2$, $\alpha = 0.05$ (two-sided), and $1 - \beta = 0.9$, the two sample binary outcome formula yields a sample size of $N = 716$ cases and $N = 716$ controls

Case-Control Sample Size

- Cases are often harder to obtain than controls
- How many controls per case?
- Continuous exposure model
- Discrete exposure model

Case-Control: Continuous Exposure

- Ury (Biometrics 1975)

- Cases

$$Y_{1i} = \mu_i + \delta + \epsilon_{1i}; \quad i = 1, \dots, N$$

- Controls (k for each case)

$$Y_{2ij} = \mu_i + \epsilon_{2ij}; \quad j = 1, \dots, k$$

- Assume $\epsilon_{1i}, \epsilon_{2ij}$ iid with

$$E(\epsilon_{1i}) = E(\epsilon_{2ij}) = 0$$

$$\text{Var}(\epsilon_{1i}) = \text{Var}(\epsilon_{2ij}) = \sigma^2$$

Case-Control: Continuous Exposure

- Let

$$\bar{Y}_{2i} = \frac{1}{k} \sum_{j=1}^k Y_{2ij}$$

- Then a consistent and unbiased estimator of the exposure effect is

$$\hat{\delta}_k = \frac{1}{N} \sum_{i=1}^N (Y_{1i} - \bar{Y}_{2i}) \equiv \bar{Y}_1 - \bar{Y}_2$$

Case-Control: Continuous Exposure

- By independence and homogeneity of variance assumptions

$$\text{Var}(\bar{Y}_1) = \frac{\sigma^2}{N} \quad \text{and} \quad \text{Var}(\bar{Y}_2) = \frac{\sigma^2}{kN}$$

- Therefore

$$\text{Var}(\hat{\delta}_k) = \frac{\sigma^2}{N} \left(\frac{k+1}{k} \right)$$

- For $k = 1$,

$$\text{Var}(\hat{\delta}_1) = \frac{2\sigma^2}{N}$$

Case-Control: Continuous Exposure

- Relative efficiency

$$\text{eff}(\hat{\delta}_1, \hat{\delta}_k) = \frac{\text{Var}(\hat{\delta}_k)}{\text{Var}(\hat{\delta}_1)} = \frac{k+1}{2k} \rightarrow \frac{1}{2} \text{ as } k \rightarrow \infty$$

k	$\text{eff}(\hat{\delta}_1, \hat{\delta}_k)$
1	1.00
2	0.75
3	0.67
4	0.63
5	0.60
10	0.55
∞	0.50

Case-Control: Continuous Exposure

- Assuming N large or $\epsilon_{1i}, \epsilon_{2ij} \sim N(0, \sigma^2)$
- Under $H_0 : \delta = 0$

$$Z = \frac{\hat{\delta}_k}{\sqrt{\text{Var}(\hat{\delta}_k)}} \sim N(0, 1)$$

- Under $H_A : \delta = \delta_A > 0$,

$$1 - \beta = \Pr \left[\frac{\hat{\delta}_k - \delta_A}{\sqrt{\text{Var}(\hat{\delta}_k)}} > z_{1-\alpha/2} - \frac{\delta_A}{\sqrt{\text{Var}(\hat{\delta}_k)}} \right]$$

Case-Control: Continuous Exposure

- Implying

$$-z_{1-\beta} = z_{1-\alpha/2} - \frac{\delta_A}{\sqrt{\text{Var}(\hat{\delta}_k)}}$$

$$(z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{\delta_A^2}{\text{Var}(\hat{\delta}_k)} = \delta_A^2 \frac{Nk}{\sigma^2(k+1)}$$

$$N = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2(k+1)}{\delta_A^2} \frac{2k}{2k}$$

Case-Control: Continuous Exposure

- So, for the two sample problem, compute the usual sample size N per arm assuming an equal sample size per arm
- Multiply N by $(k + 1)/(2k)$ to get the number of cases
- Multiply N by $(k + 1)/2$ to get the number of controls

Case-Control: Discrete Exposure

- The same relative efficiency result holds (Ury, Biometrics 1975)
- Here comparing 1:1 vs. $k:1$ controls to cases using a generalization of McNemar/MH
- Same sample size computation; cf. Note 17.2 in the text

Case-Control, Discrete Exposure: Example

- Suppose that with one control per case, we calculate that 716 cases and 716 controls are needed to achieve a particular α and β
- Then with 2 controls per case, we need $716 \times 3/4 = 537$ cases and 1074 controls

Outline

- Two sample: Continuous
- Two sample: Binary
- Case-control studies
- Estimating power with simulations

Power and Sample Size

- Determination of power / sample size is important for many reasons
- Under-powered: May miss scientifically meaningful differences
- Over-powered: Waste of resources
- Ethics
- How can one compute power / sample size in more complicated situations than those addressed in the notes or text? For example, what is the power of the Kruskal-Wallis test for a fixed sample size?

Sample Size Calculation by Simulation

- One approach: Conduct a simulation study
 1. Simulate a single data set of size N under a particular alternative
 2. Evaluate test statistic for the simulated data set; record whether reject H_0
 3. Repeat steps 1 and 2 multiple times (e.g., 10,000)
 4. Compute the proportion of simulated data sets for which H_0 is rejected; this is an estimate of power
 5. If the estimated power is larger than required, reduce N and repeat steps 1-4; if the estimated power is too low, increase N and repeat steps 1-4

Simulated Power

- To help determine if the simulation is working correctly, check the following:
 - Simulate data sets under the null. Then the proportion of simulated data sets for which H_0 is rejected should approximate the specified type I error rate α
 - As one moves away from H_0 , the estimated power should increase towards 1
- In step 3 on the previous page, use a relatively small number of simulated datasets until close to the desired power, then increase the number of datasets to obtain a more accurate estimate

Two Sample Test: σ Unknown

- Recall

```
> power.t.test(59, delta=15, sd=25)

Two-sample t test power calculation

      n = 59
      delta = 15
        sd = 25
   sig.level = 0.05
      power = 0.8982732
    alternative = two.sided

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

- Let's run a simulation and compare the estimated power to this result

Simulated Power Using R

```
set.seed(251); n <- 59; sd <- 25
mysim <- function(mdiff,nsims){
    rejects <- 0
    for (ii in 1:nsims){
        y1 <- rnorm(n,0,sd)
        y2 <- rnorm(n,mdiff,sd)
        tt <- t.test(y1,y2,var.equal=T)
        if (tt$p.value<0.05) rejects <- rejects + 1
    }
    print(paste("mdiff:",mdiff,", estimated power:",rejects/nsims))
}
mysim(0,10000)
mysim(10,100)
mysim(10,100)
mysim(15,10000)
mysim(20,1000)

[1] "mdiff: 0 , estimated power: 0.0505"
[1] "mdiff: 10 , estimated power: 0.54"
[1] "mdiff: 10 , estimated power: 0.51"
[1] "mdiff: 15 , estimated power: 0.9019"
[1] "mdiff: 20 , estimated power: 0.993"
```

Simulated Power Using SAS

```
%macro epower(mdif=,seed=);

%let i=1;    %let n=59;    %let sd=25;    %let nsims=10000;

data;
  %do i = 1 %to &nsims;
    i=&i;
    do j=1 to &n; y=rannor(&seed)*&sd; group=1; output; end;
    do j=1 to &n; y=rannor(&seed)*&sd + &mdif; group=2; output; end;
  %end;

ods output ttests=ttests;
proc ttest; class group; var y; by i; run;
data ttests; set ttests;
  if method="Pooled";
  reject=0; if Probt<0.05 then reject=1;

proc freq data=ttests; table reject; run;
%mend;

%epower(mdif=15,seed=97231);
```

Simulated Power Using SAS cont.

- Reason for:

```
if method="Pooled";
```

- Here's the dataset produced when $i = 1$:

Obs	i	Variable	Method	Variances	tValue	DF	Probt
1	1	y	Pooled	Equal	-3.51	116	0.0006
2	1	y	Satterthwaite	Unequal	-3.51	115.86	0.0006

- Output of the simulation;

The FREQ Procedure				
reject	Frequency	Percent	Cumulative	Cumulative
			Frequency	Percent
0	1054	10.54	1054	10.54
1	8946	89.46	10000	100.00

BIOS 662 Fall 2018

Power and Sample Size, Part III

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Power and Sample Size

- Many of the sample size/power formulas assume a balanced design (exception: case-control)
- How do we generalize to unbalanced designs?
- For example, consider a two-sample t test with a continuous outcome

Two Sample t Test

- Assume normality and homogeneity of variance

$$\bar{Y}_i \sim N(\mu_i, \sigma^2/N_i) \text{ for } i = 1, 2$$

- Under $H_0 : \mu_1 - \mu_2 = 0$

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{1/N_1 + 1/N_2}} \sim t_{N_1+N_2-2}$$

- Under $H_A : \mu_1 - \mu_2 = \delta_A > 0$

$$\bar{Y}_1 - \bar{Y}_2 \sim N(\delta_A, \sigma^2(1/N_1 + 1/N_2))$$

implying

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{1/N_1 + 1/N_2}} \sim N\left(\frac{\delta_A}{\sigma \sqrt{1/N_1 + 1/N_2}}, 1\right)$$

Two Sample t Test

- Note

$$\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{\sigma^2} \sim \chi_{N_1+N_2-2}^2$$

- Therefore

$$\frac{(\bar{Y}_1 - \bar{Y}_2)/(\sigma \sqrt{1/N_1 + 1/N_2})}{\sqrt{\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{\sigma^2(N_1+N_2-2)}}} \sim t_{N_1+N_2-2, \delta_A/(\sigma \sqrt{1/N_1+1/N_2})}$$

- Equivalently

$$\frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{1/N_1 + 1/N_2}} \sim t_{N_1+N_2-2, \delta_A/(\sigma \sqrt{1/N_1+1/N_2})}$$

Two Sample t Test

- Given N_1 , N_2 , α and δ_A , power is $\Pr[T > t_{1-\alpha/2}]$ where

$$T \sim t_{N_1+N_2-2, \delta_A/(\sigma\sqrt{1/N_1+1/N_2})}$$

- For example, suppose $N_1 = 10$, $N_2 = 20$, $\alpha = 0.05$,
 $\delta_A = 15$ and $\sigma = 25$
- R

```
> 1-pt(qt(0.975,28),28,15/(25*sqrt(1/10+1/20)))
[1] 0.3214083
```

Two Sample t Test

- SAS

```
proc power;  
    twosamplemeans  
    meandiff = 15  
    groupns = 10|20  
    stddev = 25  
    power = .;
```

Two-sample t Test for Mean Difference

Fixed Scenario Elements

Distribution	Normal
Method	Exact
Mean Difference	15
Standard Deviation	25
Group 1 Sample Size	10
Group 2 Sample Size	20

Power

0.322

Drop-outs and Loss to Follow-up

- Drop-out:
 - one who terminates involvement in an activity
 - a subject who withdraws from a trial or follow-up study by an announced unwillingness to continue to submit to required procedures
 - a subject who refuses or stops taking the assigned treatment
 - a subject who misses a scheduled visit
- A drop-out may drop out of treatment and/or out of follow-up

Drop-outs and Loss to Follow-up

- Drop-in:
 - a subject enrolled in a clinical trial who receives a study treatment different from or in addition to the assigned treatment
- Lost to follow-up:
 - a subject who cannot be followed for some outcome or observation of interest
- In terms of its potential effect on the *validity* of the study, loss to follow-up is a more critical issue than dropping out of or into treatment groups
- Analyze on basis of the intention-to-treat principle

Adjusting the Sample Size

- For loss to follow-up, can scale up N by the proportion of subjects one anticipates may be lost to follow-up
- For drop-in and drop-out of treatment groups, make an allowance for the estimated (guessed) proportions of subjects dropping in and out and then adjust Δ
- Example: Clinical trial treating alcohol dependence
 - Outcome measure: Percent days abstinent (PDA)
 - Effective treatment assumed to increase mean PDA by 10 percentage points
 - Allowing for 25% drop-out of treatment, net effect is a mean increase of 7.5 percentage points

Adjusting the Sample Size: Example

- A clinical trial was planned to investigate reducing the risk of new cardiovascular events in subjects with history of CVD and periodontal disease
- Treatment groups: Study-supplied periodontal therapy versus usual care from own dentist
- Outcome measure: CVD event rate
- Effective treatment assumed to reduce rate by 25%
- Assume event rate of 6.5% p.a. in usual care group
- So assumed rate in active group is 4.875% p.a.

Adjusting the Sample Size: Example cont.

- Assume 10% in active care group drop out of treatment and 5% in usual care group drop in
- Net event rate in usual care group:
 - 95% have rate 6.5%; 5% have rate 4.875%
 - net rate: $0.95 \times 6.5\% + 0.05 \times 4.875\% = 6.42\%$
- Net event rate in active treatment group:
 - 90% have rate 4.875%; 10% have rate 6.5%
 - net rate: $0.9 \times 4.875\% + 0.1 \times 6.5\% = 5.04\%$
- Use the net rates in the power / sample size calculations

Correlation Coefficient / Linear Regression

- Let ρ be the correlation between X and Y and r be the sample correlation based on n pairs of observations
- Fisher's transform is approximately normally distributed:

$$Z_r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right) \sim N \left(\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{N-3} \right)$$

- For the test $H_0 : \rho = \rho_0$ versus $H_A : \rho \neq \rho_0$ to have power $1 - \beta$ for the specific alternative $\rho = \rho_1$, we need a sample of size

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(Z_{\rho_1} - Z_{\rho_0})^2} + 3$$

Correlation Coefficient / Linear Regression

- For a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

we have

$$\hat{\beta}_1 = \frac{s_Y}{s_X} r$$

or

$$r = \frac{s_X}{s_Y} \hat{\beta}_1$$

where s_X and s_Y are the sample standard deviations of X and Y

Linear Regression Example

- Suppose we want to find the appropriate sample size to have 90% power to detect $\beta_1 = 0.5$ and from previous studies we have estimates $s_X = 2$ and $s_Y = 10$
- If $\widehat{\beta}_1 = 0.5$ then

$$r = \frac{s_X}{s_Y} \widehat{\beta}_1 = \frac{2}{10} \cdot 0.5 = 0.1$$

- We should have the same power to test

$$H_A : \beta_1 = 0.5 \text{ against } H_0 : \beta_1 = 0$$

as to test

$$H_A : \rho = 0.1 \text{ against } H_0 : \rho = 0$$

Linear Regression Example

- Here

$$Z_0 = \frac{1}{2} \log \left(\frac{1+0}{1-0} \right) = 0$$

and

$$Z_1 = \frac{1}{2} \log \left(\frac{1+0.1}{1-0.1} \right) = 0.1003$$

- The sample size n to give us power $1 - \beta$ for testing $\rho = 0.1$ versus $\rho = 0$ is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(Z_{\rho_1} - Z_{\rho_0})^2} + 3$$

$$= \frac{(1.96 + 1.28)^2}{(0.1003 - 0)^2} + 3$$

$$= 1046$$

Adjusting for Covariates in Regression Models

Reference: Hsieh, Bloch, Larsen (1998) A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* 17:1623-1634.

- Suppose that our primary exposure of interest is X_1 and we want to adjust for covariates, X_2, X_3, X_4, \dots
- Calculate R^2 from a regression model of X_1 (not Y) as a function of the other covariates
- Adjust the sample size using a variance inflation factor (VIF):

$$\text{VIF} = \frac{1}{1 - R^2}$$

and use sample size $n' = \text{VIF} \cdot n$ where n is the sample size for a simple linear regression of Y on X_1

Example Adjusting for Covariates

- Continuing the previous example, suppose we want to adjust for just one additional variable X_2 , with our interest still in the sample size to give 90% power to detect $\beta_1 = 0.5$
- Suppose the correlation between X_1 and X_2 is $r = 0.3$
- Then

$$\begin{aligned} n' &= \text{VIF} \cdot n = \frac{1}{1 - R^2} \cdot n = \frac{1}{1 - r^2} \cdot n \\ &= \frac{1}{1 - 0.3^2} \cdot 1046 = 1149 \end{aligned}$$

BIOS 662 Fall 2018

Clustered Data

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Correlated Data

- To this point, all methods have assumed data are iid
- What do we do if dependencies exist between observations?
- Typically, correlated data occur in clusters/groups
- Examples:
 - Repeated measures on individuals over time
 - Natural groupings of individuals (e.g., litters, schools)
- Can occur in observational or randomized studies;
an example of the latter is a cluster randomized study

Cluster Randomized Studies

- Also known as *group allocation* designs
- Section 18.4 of the text (deals with correlation structures)
- Suppose we want to compare two school-based methods of smoking prevention in teenagers
- We may randomly assign interventions to schools, but measure smoking in children

Central Issue

- How do we do testing, estimation, sample size calculations, etc., taking into account that responses within a cluster/group (e.g., school) may not be independent?
- Use methods allowing for dependency (correlation) within groups but assuming independence (no correlation) between groups

Continuous Response Model

- Let Y_{ijk} = response of the k^{th} person in the j^{th} cluster at the i^{th} treatment level,

$$i = 1, 2, \dots, I; \quad j = 1, 2, \dots, J; \quad k = 1, 2, \dots, K$$

- Let

$$\bar{Y}_{ij} = \frac{1}{K} \sum_{k=1}^K Y_{ijk}$$

- Assume:

$$E(Y_{ijk}) = \mu_i; \quad \text{Var}(Y_{ijk}) = \sigma^2$$

$$\text{Cov}(Y_{ijk}, Y_{ijk'}) = \rho\sigma^2; \quad \text{Cov}(Y_{ijk}, Y_{ij'k'}) = 0$$

Continuous Response Model

- Then

$$\text{Var}(\bar{Y}_{ij}) = E(\bar{Y}_{ij}^2) - \mu_i^2$$

$$= K^{-2} E\left(\sum_{k=1}^K Y_{ijk}\right)^2 - \mu_i^2$$

$$= K^{-2} E\left(\sum_{k=1}^K Y_{ijk}^2 + \sum_{k \neq k'} \sum Y_{ijk} Y_{ijk'}\right) - \mu_i^2$$

$$= K^{-2} \left(K\sigma^2 + K(K-1)\rho\sigma^2 \right)$$

$$= \frac{\sigma^2}{K} \left(1 + (K-1)\rho \right)$$

Variance Inflation Factor (VIF)

- $(1 + (K - 1)\rho)$ is the *variance inflation factor* (VIF)
- It measures the increase in the variance of the mean due to the within-subject correlation of measurements (ρ)
- $\text{VIF} > 1$ for $\rho > 0$ and $K > 1$
- Let

$$\bar{Y}_i = \frac{\sum_j \bar{Y}_{ij}}{J} = \frac{\sum_{j,k} Y_{ijk}}{JK}$$

- Then

$$\text{Var}(\bar{Y}_i) = \frac{\sigma^2}{JK} \text{VIF}$$

Continuous Response Model

- Suppose $I = 2$ and $n_1 = n_2 = JK$
- If we ignore the correlation within cluster

$$z_{\text{ignore}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{1/n_1 + 1/n_2}},$$

- Should instead use

$$\begin{aligned} z_{\text{true}} &= \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma \sqrt{(1/n_1 + 1/n_2) \cdot \text{VIF}}} \\ &= \frac{z_{\text{ignore}}}{\sqrt{\text{VIF}}} \end{aligned}$$

Effect of Correlation

- $|z_{\text{true}}| < |z_{\text{ignore}}|$ for $\rho > 0$ and $K > 1$
- Thus ignoring correlation will lead to inflated type I error
- Intuition: Naïve approach acts as if we have more information than we do

Sample Size When $I = 2$

- Sample size per arm

$$n = 2 \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta} \right)^2 \text{VIF}$$

where $\Delta = |\mu_1 - \mu_2|/\sigma$

- If $\rho = 0$, then $\text{VIF} = 1$ and

$$n = 2 \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta} \right)^2$$

- If $\rho = 1$, then $\text{VIF} = K$

$$n = 2 \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta} \right)^2 \cdot K$$

- Typically $0.1 \leq \rho \leq 0.4$

Variance Inflation Factors

- Table 18.4 in the text:

K	ρ				
	0.001	0.01	0.02	0.05	0.1
2	1.001	1.01	1.02	1.05	1.10
5	1.004	1.04	1.09	1.20	1.40
10	1.009	1.09	1.18	1.45	1.90
100	1.099	1.99	2.98	5.95	10.90
1000	1.999	10.99	20.98	50.95	100.90

Concluding Remarks

- What if cluster/group sizes vary? Say $k = 1, \dots, K_j$
- Use expected cluster size; cf. Manatunga, Hudgens, Chen (*Biometrical Journal*, 2001)
- Methods for analyzing clustered data include mixed models and generalized estimating equations (BIOS 762/3/7)

BIOS 662 Fall 2018

Rates and Proportions

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Prevalence/incidence
- Direct standardization
- Indirect standardization

Rates and Proportions

- Cf. chapter 15 of the text
- *Prevalence*: Proportion (π) of people with a particular disease at a fixed point in time
- *Rate*: Change in a variable over a specified time interval divided by the length of the time interval
- *Incidence*: The number of new cases of a disease in a period of time divided by the person-years at risk
- Incidence is a rate, prevalence is not

Prevalence

- Consider a random sample of size N from the population of interest
- Suppose n have the disease of interest (“cases”)
- Estimator of prevalence:

$$\hat{p} = \frac{n}{N} = \frac{\text{number of cases}}{\text{sample size}}$$

- CIs and tests for prevalence are based on

$$n \sim \text{Binomial}(N, \pi)$$

where π is the population prevalence

Prevalence: Example

- A random sample of 1,717 injecting drug users in 6 major cities in the U.S. found that 206 were HIV positive.
- Estimated prevalence of HIV among injecting drug users

$$\hat{p} = 206/1717 = 0.120$$

- Large sample 95 % CI: (0.105, 0.135)

Incidence

- Estimator of incidence:

$$\hat{I} = \frac{\text{number of new cases}}{(\text{sample size}) \times (\text{time interval})}$$

(This is a simplified version; if we use the definition of incidence strictly, we should exclude time after a person has developed the disease)

- Example: Incidence of diabetes among Pima Indians

- $N = 1,728$, time = 6 years, new cases = 346

[Reference: *AJE* Oct 1, 2003, page 669]

$$\hat{I} = \frac{346}{1728 \times 6} = 0.033$$

- Thus estimated incidence is 0.033 cases per person-year

Incidence

- We usually multiply by some number, such as 1,000

$$\hat{I}_{1000} = 33.4$$

- Interpretation: estimated incidence is:

33.4 cases per year per 1,000 persons

or

33.4 cases per 1,000 person years

Incidence

- Note general form

$$\hat{I}_{1000} = c \cdot \frac{\text{number of new cases}}{\text{sample size}}$$

where $c = 1000/\text{time interval}$

- Because

$$\frac{\text{number of new cases}}{\text{sample size}}$$

is a proportion, we can again use binomial principles for CIs and tests

Incidence

- Let

$$\hat{p} = \frac{n}{N} = \frac{\text{number of new cases}}{\text{sample size}}$$

so that

$$n \sim \text{Binomial}(N, \pi)$$

- Note that the π here is distinct from in the prevalence situation; now it is the probability of becoming a case in the follow-up interval

- Thus

$$\widehat{\text{Var}}(\hat{p}) = \hat{p}(1 - \hat{p})/N$$

implying

$$\widehat{\text{Var}}(\hat{I}_{1000}) = c^2 \hat{p}(1 - \hat{p})/N$$

Incidence CI

- Approximate $100(1 - \alpha)\%$ CI

$$\hat{I}_{1000} \pm z_{1-\alpha/2} \sqrt{c^2 \hat{p}(1 - \hat{p})/N}$$

- Diabetes example:

$$\hat{p} = \frac{346}{1728} = 0.20; \quad c = \frac{1000}{6} = 166.67$$

- 95% CI

$$33.4 \pm 3.14 = (30.2, 36.5)$$

Direct Standardization

- We may need to adjust rates/proportions for possible confounders, e.g., age, gender
- Example: Study of smoking in China (1984)
 - Urban women: 1,320 questioned, 330 current smokers
 - Rural women: 1,338 questioned, 414 current smokers
$$\hat{p}_u = 330/1320 = 0.25; \quad \hat{p}_r = 414/1338 = 0.31$$
- Concern: Age may be a confounder

Direct Standardization

- Three steps
 1. Divide samples into K categories of the potential confounder
 2. Compute the proportion or rate in each confounder category
 3. Compute the weighted average of confounder-specific proportions/rates
- Choice of weights is based on a *standard or reference population*; e.g., aggregate of samples in hand, governmental population survey

Direct Standardization

- China smoking example

Age	Urban			Rural		
	N_{1k}	n_{1k}	\hat{p}_{1k}	N_{2k}	n_{2k}	\hat{p}_{2k}
35-39	129	8	0.062	387	44	0.114
40-44	243	53	0.218	441	138	0.313
45-49	478	135	0.282	300	130	0.433
50-54	470	134	0.285	210	102	0.486

Direct Standardization

- Combined age distribution

Age	N_k	w_k
35-39	516	0.194
40-44	684	0.257
45-49	778	0.293
50-54	680	0.256
Total	2658	1.000

Direct Standardization

- Adjusted prevalence estimator

$$\hat{p}_{j_{\text{adj}}} = \frac{\sum_{k=1}^K w_k \hat{p}_{jk}}{\sum_{k=1}^K w_k}$$

- Estimator of prevalence in the reference (i.e., standard) population is based on the observed rates from the study population

Direct Standardization

- Example: (Urban=1, Rural=2)

$$\hat{p}_{1\text{adj}} = (0.194 \times 0.062 + \dots + 0.256 \times 0.285)/1 = 0.224$$

$$\hat{p}_{2\text{adj}} = 0.354$$

- Crude difference, ratio:

$$\hat{p}_1 - \hat{p}_2 = 0.25 - 0.31 = -0.06$$

$$\hat{p}_2/\hat{p}_1 = 0.31/0.25 = 1.24$$

- Age adjusted difference, ratio:

$$\hat{p}_{1\text{adj}} - \hat{p}_{2\text{adj}} = 0.224 - 0.354 = -0.13$$

$$\hat{p}_{2\text{adj}}/\hat{p}_{1\text{adj}} = 0.354/0.224 = 1.58$$

Direct Standardization

- World Health Organization Standard Weights

Age	w_i	Age	w_i
<1	2.4	45-49	6
1-4	9.6	50-54	5
5-9	10	55-59	4
10-14	9	60-64	4
15-19	9	65-69	3
20-24	8	70-74	2
25-29	8	75-79	1
30-34	6	80-84	0.5
35-39	6	>84	0.5
40-44	6		

Direct Standardization

- China-smoking example using WHO standard:

$$\hat{p}_{1\text{adj}} = \frac{6 \times 0.062 + \dots + 5 \times 0.285}{6 + 6 + 6 + 5} = 0.209$$

$$\hat{p}_{2\text{adj}} = 0.330$$

$$\hat{p}_{1\text{adj}} - \hat{p}_{2\text{adj}} = -0.121$$

$$\frac{\hat{p}_{2\text{adj}}}{\hat{p}_{1\text{adj}}} = 1.58$$

Direct Standardization

	Crude	Combined	WHO
Difference	-0.06	-0.13	-0.12
Ratio	1.24	1.58	1.58

- Note: Combined and WHO estimates are further from null than the crude estimates
- The confounder, age, partially masks difference in smoking between urban and rural
- Intuition: Rural, older people smoke more; urban sample has greater proportion of older people

Direct Standardization

- $\hat{p}_{j\text{adj}}$ is a weighted average of independent random variables (the \hat{p}_{jk})
- Because $n_{jk} \sim \text{Binomial}(N_{jk}, \pi_{jk})$, we know that

$$\text{Var}(\hat{p}_{jk}) = \pi_{jk}(1 - \pi_{jk})/N_{jk}$$

and

$$\widehat{\text{Var}}(\hat{p}_{jk}) = \hat{p}_{jk}(1 - \hat{p}_{jk})/N_{jk}$$

Direct Standardization

- Thus

$$\widehat{\text{Var}}(\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}}) = \frac{\sum_{k=1}^K w_k^2 (\widehat{\text{Var}}(\hat{p}_{1k}) + \widehat{\text{Var}}(\hat{p}_{2k}))}{(\sum_{k=1}^K w_k)^2}$$

- Large sample tests and CIs are obtained from the CLT

Direct Standardization

- Revisiting the smoking example (using combined weights)

$$\widehat{\text{Var}}(\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}}) = 0.000318$$

- Testing $H_0 : \pi_{1_{\text{adj}}} = \pi_{2_{\text{adj}}}$ versus $H_A : \pi_{1_{\text{adj}}} \neq \pi_{2_{\text{adj}}}$,

$$Z = \frac{\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}}}{\sqrt{\widehat{\text{Var}}(\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}})}} = \frac{-0.13}{\sqrt{0.000318}} = -7.28$$

- Conclude that there is a significant difference in the prevalence of smoking between rural and urban women after adjusting for age

Standardization

- *Direct standardization:* Estimate rate or proportion in the reference population using the observed rate or proportion from the study sample
- *Indirect standardization:* Estimate the rate or proportion in the study population using the rate or proportion from the reference population

Indirect Standardization

- Suppose we observe stratum-specific prevalences (or incidences)

- Reference population: m_k/M_k for $k = 1, \dots, K$
 - Study population: n_k/N_k for $k = 1, \dots, K$

- Observed prevalence in the study population

$$\hat{p}_{\text{study}} = \frac{\sum_{k=1}^K n_k}{\sum_{k=1}^K N_k}$$

- Expected prevalence in the study population assuming stratum-specific prevalences from the reference population

$$\hat{p}_{\text{ref}} = \frac{\sum_{k=1}^K N_k m_k / M_k}{\sum_{k=1}^K N_k}$$

Indirect Standardization

- *Standardized mortality ratio* (SMR)

$$s = \frac{\hat{p}_{\text{study}}}{\hat{p}_{\text{ref}}} = \frac{\sum_{k=1}^K n_k}{\sum_{k=1}^K N_k m_k / M_k} = \frac{O}{E}$$

- Note: Calculation of s requires knowing just $\sum_k n_k$ for the study population, that is, we do not need to know the number of events for each level of the confounder
- *Standardized incidence ratio* (SIR) is defined analogously

Indirect Standardization

- The variance of s can be estimated by

$$\widehat{\text{Var}}(s) = \frac{\widehat{\text{Var}}(O) + s^2 \widehat{\text{Var}}(E)}{E^2}$$

where $\widehat{\text{Var}}(O) = \sum_k n_k$

and $\widehat{\text{Var}}(E) = \sum_k \left(\frac{N_k}{M_k} \right)^2 m_k$

- To test $H_0 : \pi_{\text{study}}/\pi_{\text{ref}} = 1$ vs. $H_0 : \pi_{\text{study}}/\pi_{\text{ref}} \neq 1$,

$$Z = \frac{s - 1}{\sqrt{\widehat{\text{Var}}(s)}} \sim N(0, 1)$$

Indirect Standardization

- Revisit smoking example
- Let's compute standardized prevalence ratio for rural women using urban women as the reference population, adjusting for age
- For rural women $O = 414$,

$$E = \frac{387 \times 8}{129} + \frac{441 \times 53}{243} + \frac{300 \times 135}{478} + \frac{210 \times 134}{470}$$
$$= 264.79$$

- Therefore $s = 414/264.79 = 1.56$

Indirect Standardization

- Now $\widehat{\text{Var}}(O) = O = 414$ and

$$\begin{aligned}\widehat{\text{Var}}(E) &= 8\left(\frac{387}{129}\right)^2 + 53\left(\frac{441}{243}\right)^2 + 135\left(\frac{300}{478}\right)^2 + 134\left(\frac{210}{470}\right)^2 \\ &= 326.49\end{aligned}$$

- Therefore

$$\begin{aligned}\widehat{\text{Var}}(s) &= \frac{\widehat{\text{Var}}(O) + s^2 \widehat{\text{Var}}(E)}{E^2} = \frac{414 + 1.56^2(326.49)}{264.79^2} \\ &= 0.0173\end{aligned}$$

implying

$$Z = \frac{s - 1}{\sqrt{\widehat{\text{Var}}(s)}} = 4.29$$

Indirect Standardization

- When computing standardized rates or proportions, inspect observed and expected cells (if feasible) to facilitate understanding

Age	O_k	E_k	O_k/E_k
35-39	44	24.0	1.83
40-44	138	96.2	1.43
45-49	130	84.7	1.53
50-55	102	59.9	1.70

BIOS 662 Fall 2016

Survival Analysis

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Introduction to survival data/analysis
- Kaplan-Meier estimator, standard error and CI
- Log-rank test
- (Cox / proportional hazards model)

Survival Analysis

- Chapter 16 of the text; BIOS 680/780
- Survival analysis: Response is time to an event
- Measure time from beginning of follow-up until an event such as incident disease, death, or relapse
- In a clinical trial, the beginning of follow-up is almost always the time of randomization
- In an epidemiology study, beginning of follow-up is usually the time of (initial) exposure assessment
- Examples:
 - time from kidney transplant until death
 - time from leukemia treatment to remission

Survival Analysis: Notation

- Let T^* denote the (possibly unknown) survival time;
assume $T^* > 0$
- Define the survival function

$$S(t) = \Pr[T^* > t] = 1 - \Pr[T^* \leq t] = 1 - F(t)$$

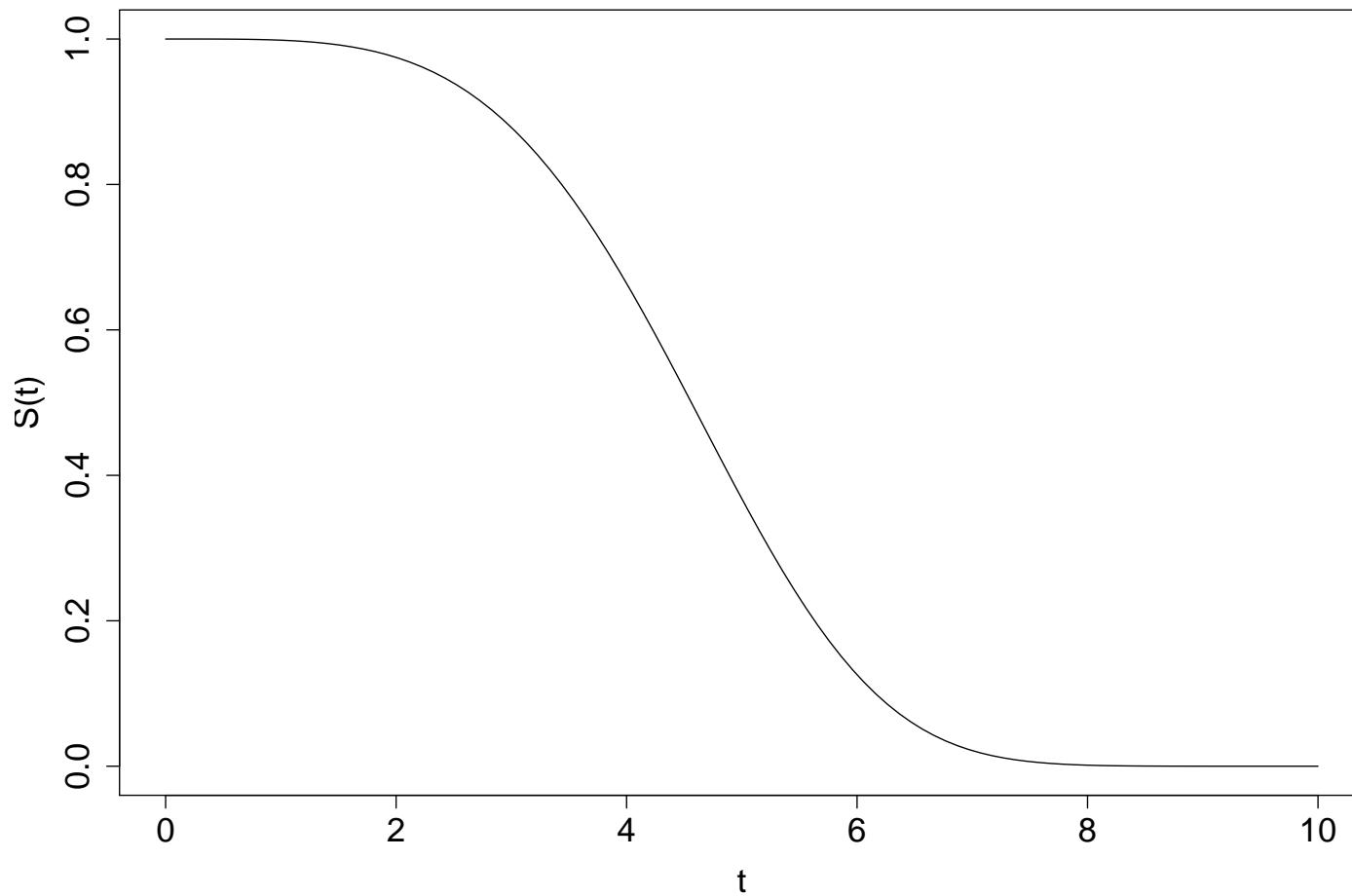
where $F(t)$ is the CDF of T^*

- Properties:

$$S(0) = 1; \quad S(\infty) = 0$$

If $t_1 \leq t_2$, then $S(t_1) \geq S(t_2)$

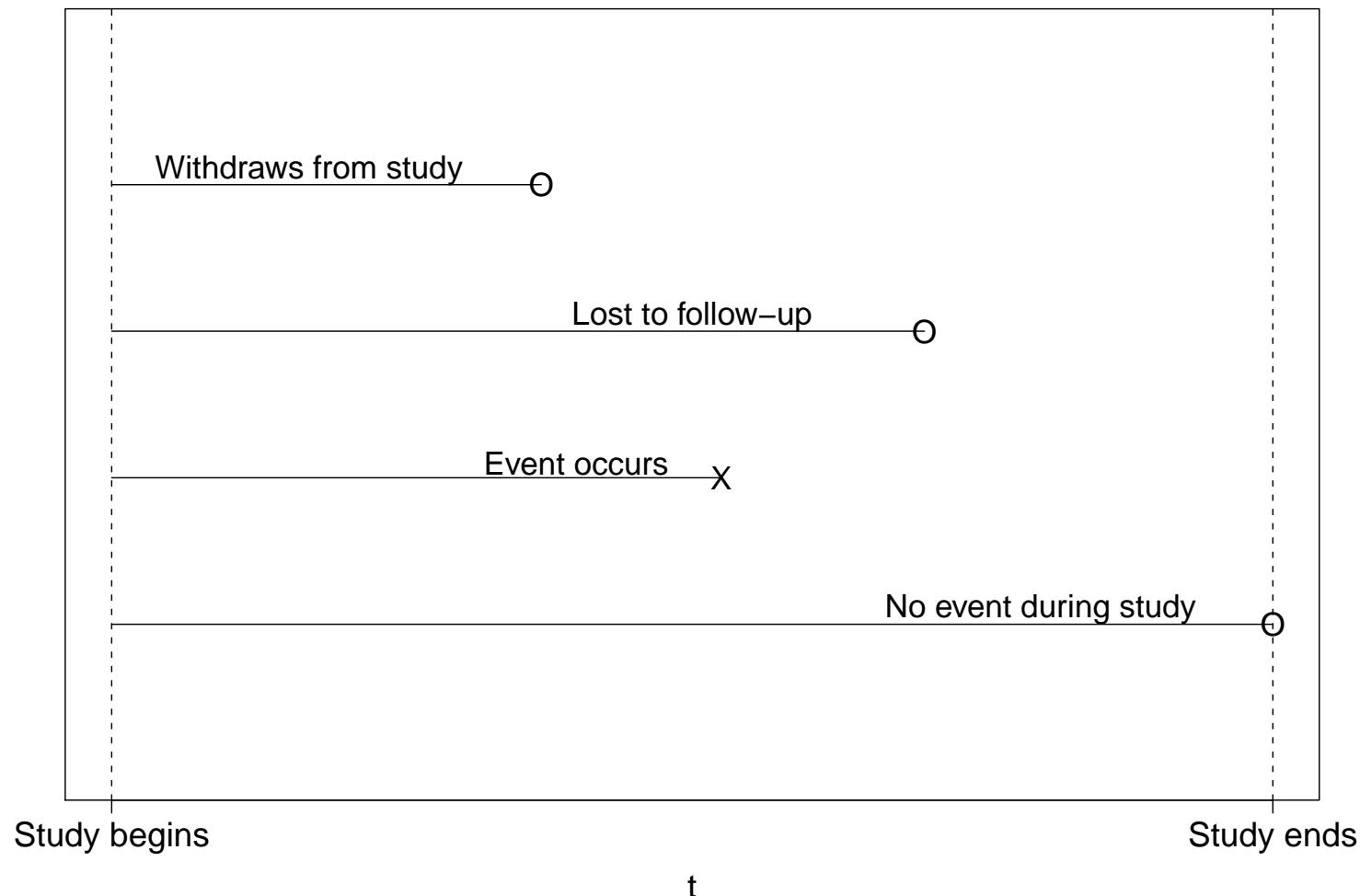
Example Survival Curve/Function



Censoring

- Often we do not know the exact time of failure of all subjects
- Reasons for **right** censoring:
 - subject does not experience the event of interest before the end of the study
 - subject is lost to follow-up during the study (e.g., withdraws from study, moves, dies from something other than the event of interest)
- Failure times can also be left or interval censored

Right Censoring



Survival Data

- Let T_i^* and C_i denote the survival and right censoring times for the i^{th} individual
- Observe $T_i = \min\{T_i^*, C_i\}$
- Censoring indicator

$$\delta_i = \begin{cases} 1 & \text{if failure, i.e., } T_i = T_i^* \\ 0 & \text{if right censored, i.e., } T_i = C_i \end{cases}$$

- We observe (T_i, δ_i) for $i = 1, 2, \dots, N$

Example

- Remission time in weeks for leukemia patients ($N = 21$)

(T_i, δ_i)	(T_i, δ_i)	(T_i, δ_i)
(6,1)	(6,1)	(6,1)
(6,0)	(7,1)	(9,0)
(10,1)	(10,0)	(11,0)
(13,1)	(16,1)	(17,0)
(19,0)	(20,0)	(22,1)
(23,1)	(25,0)	(32,0)
(32,0)	(34,0)	(35,0)

Estimation

- How do we estimate $S(t)$ with minimal assumptions?
- Answer 1: In the absence of censoring, use $1 - \text{EDF}$
- Answer 2: Otherwise, use the Kaplan-Meier estimator

Tabular Summary of Data

- Let $t_{(1)}, t_{(2)}, \dots, t_{(J)}$ be the distinct ordered failure times (censoring times are ignored)

Failure time	Risk set	No. of failures	No. censored in $[t_{(j)}, t_{(j+1)})$
$t_{(j)}$	$R(t_{(j)})$	m_j	q_j
$t_{(0)} = 0$	$R(t_{(0)}) = N$	$m_0 = 0$	q_0
$t_{(1)}$	$R(t_{(1)})$	m_1	q_1
$t_{(2)}$	$R(t_{(2)})$	m_2	q_2
\vdots	\vdots	\vdots	\vdots
$t_{(J)}$	$R(t_{(J)})$	m_J	q_J

- $R(t_{(j)}) = R(t_{(j-1)}) - m_{j-1} - q_{j-1}$

Leukemia Example

$t_{(j)}$	$R(t_{(j)})$	m_j	q_j
0	21	0	0
6	21	3	1
7	17	1	1
10	15	1	2
13	12	1	0
16	11	1	3
22	7	1	0
23	6	1	5

Kaplan-Meier Estimator of $S(t)$

- For $t \in [0, t_{(1)})$

$$\hat{S}(t) = 1$$

- For $t \in [t_{(j)}, t_{(j+1)})$

$$\hat{S}(t) = \hat{S}(t_{(j-1)}) \cdot \widehat{\Pr}[T > t_{(j)} | T \geq t_{(j)}]$$

$$= \hat{S}(t_{(j-1)}) \left(\frac{R(t_{(j)}) - m_j}{R(t_{(j)})} \right)$$

- Assumes anyone censored at time $t_{(j)}$ has $T > t_{(j)}$

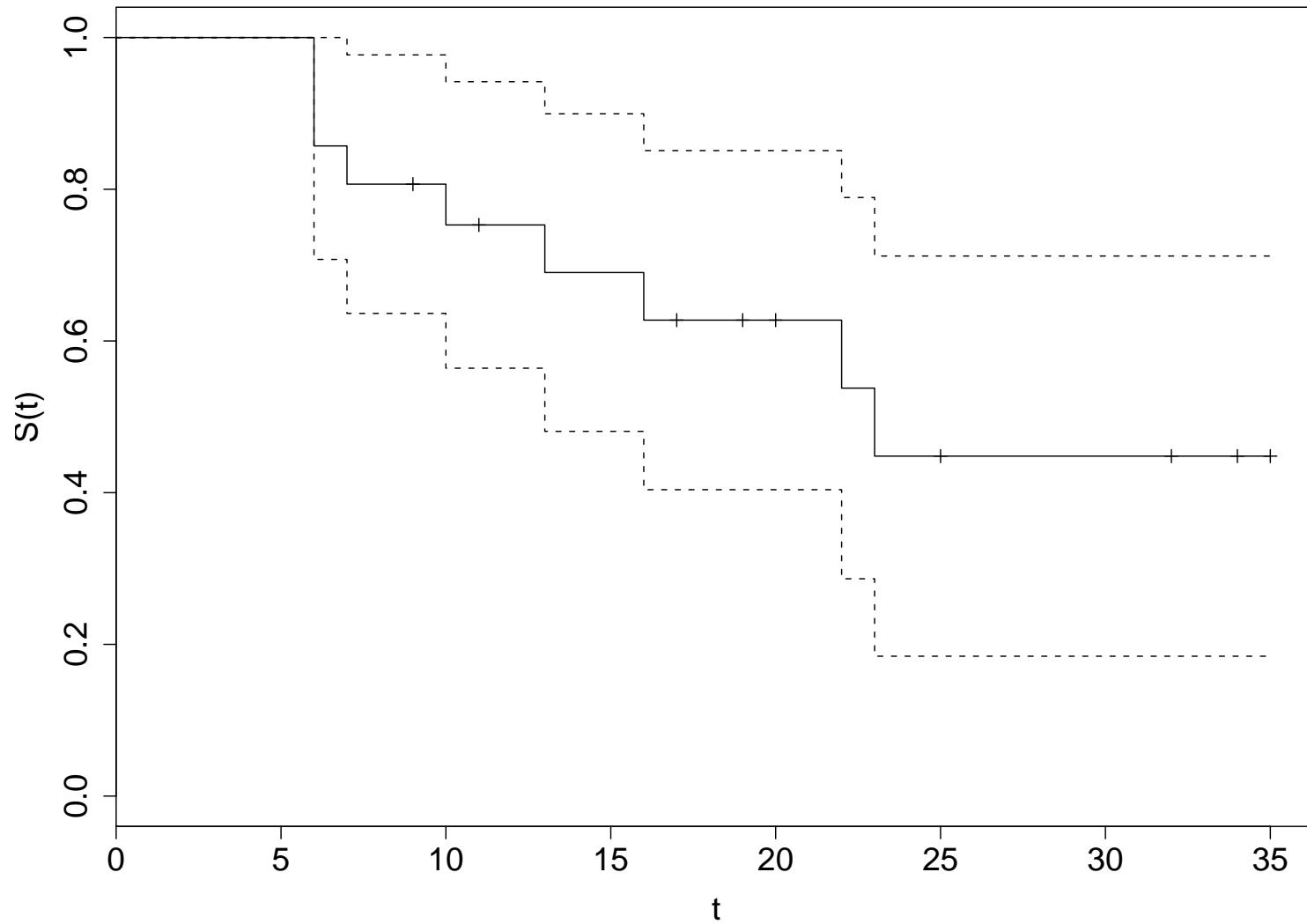
Kaplan-Meier Estimator

- KM is a nonparametric maximum likelihood estimator (NPMLE))
- Assumes independent censoring
- Also known as the *product limit estimator*
- If no censoring, KM equals $1 - \text{EDF}$
- Alternative: *Life-table or actuarial method*

Leukemia Example

$t_{(j)}$	$R(t_{(j)})$	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1
6	21	3	1	$18/21 = 0.857$
7	17	1	1	$0.857(16/17) = 0.807$
10	15	1	2	$0.807(14/15) = 0.753$
13	12	1	0	$0.753(11/12) = 0.690$
16	11	1	3	$0.690(10/11) = 0.627$
22	7	1	0	$0.627(6/7) = 0.538$
23	6	1	5	$0.538(5/6) = 0.448$

Kaplan-Meier Estimate for Leukemia Example



Kaplan-Meier Estimate: R

```
> t <- c(6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35)
> delta <- c(1,1,1,0,1,0,1,0,0,1,1,0,0,0,1,1,0,0,0,0,0,0)
> x <- rep(1,21)

> library("survival")
> fit <- survfit(Surv(t, delta)~x ,conf.type="plain")

> plot(fit,xlab="t",ylab="S(t)")

> summary(fit)
```

```
Call: survfit(formula = Surv(t, delta) ~ x, conf.type = "plain")
```

time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI
6	21	3	0.857	0.0764	0.707		1.000	
7	17	1	0.807	0.0869	0.636		0.977	
10	15	1	0.753	0.0963	0.564		0.942	
13	12	1	0.690	0.1068	0.481		0.900	
16	11	1	0.627	0.1141	0.404		0.851	
22	7	1	0.538	0.1282	0.286		0.789	
23	6	1	0.448	0.1346	0.184		0.712	

Kaplan-Meier Estimate: SAS

```
proc lifetest;  
    time t*delta(0);
```

The LIFETEST Procedure
Product-Limit Survival Estimates

t	Survival	Failure	Survival		
			Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	21
6.0000	.	.	.	1	20
6.0000	.	.	.	2	19
6.0000	0.8571	0.1429	0.0764	3	18
6.0000*	.	.	.	3	17
7.0000	0.8067	0.1933	0.0869	4	16
9.0000*	.	.	.	4	15
10.0000	0.7529	0.2471	0.0963	5	14
10.0000*	.	.	.	5	13
11.0000*	.	.	.	5	12
13.0000	0.6902	0.3098	0.1068	6	11
16.0000	0.6275	0.3725	0.1141	7	10

Kaplan-Meier Estimate: SAS cont.

17.0000*	.	.	.	7	9
19.0000*	.	.	.	7	8
20.0000*	.	.	.	7	7
22.0000	0.5378	0.4622	0.1282	8	6
23.0000	0.4482	0.5518	0.1346	9	5
25.0000*	.	.	.	9	4
32.0000*	.	.	.	9	3
32.0000*	.	.	.	9	2
34.0000*	.	.	.	9	1
35.0000*	.	.	.	9	0

NOTE: The marked survival times are censored observations.

Summary of the Number of Censored and Uncensored Values

			Percent
Total	Failed	Censored	Censored
21	9	12	57.14

Greenwood SE/CI of KM

- Let $n_j = R(t_{(j)})$
- Write the Kaplan-Meier estimator as

$$\hat{S}(t) = \prod_{j=1}^i \hat{p}_j \quad \text{for } t \in [t_{(i)}, t_{(i+1)}),$$

where $\hat{p}_j = (n_j - m_j)/n_j$ is the estimated probability of surviving interval $[t_{(j)}, t_{(j+1)})$ conditional on survival up to $t_{(j)}$

Greenwood SE/CI for KM

- Take logs

$$\log \hat{S}(t) = \sum_{j=1}^i \log \hat{p}_j$$

so that

$$\text{Var}(\log \hat{S}(t)) = \sum_{j=1}^i \text{Var}(\log \hat{p}_j)$$

- Binomial argument

$$\widehat{\text{Var}}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$$

Greenwood SE/CI for KM

- Taylor series approximation

$$\widehat{\text{Var}}(g(X)) \approx (g'(\mu))^2 \widehat{\text{Var}}(X)$$

implies

$$\begin{aligned}\widehat{\text{Var}}(\log \hat{p}_j) &\approx \left(\frac{1}{\hat{p}_j}\right)^2 \left(\frac{\hat{p}_j(1 - \hat{p}_j)}{n_j}\right) = \frac{1 - \hat{p}_j}{n_j \hat{p}_j} \\ &= \frac{m_j}{n_j(n_j - m_j)}\end{aligned}$$

- Thus

$$\widehat{\text{Var}}(\log \hat{S}(t)) \approx \sum_{j=1}^i \frac{m_j}{n_j(n_j - m_j)}$$

Greenwood SE/CI for KM

- Additional application of Taylor series approximation

$$\widehat{\text{Var}}(\log \hat{S}(t)) \approx (\hat{S}(t))^{-2} \widehat{\text{Var}}(\hat{S}(t))$$

implying

$$\widehat{\text{Var}}(\hat{S}(t)) \approx (\hat{S}(t))^2 \sum_{j=1}^i \frac{m_j}{n_j(n_j - m_j)}$$

- Thus

$$\widehat{\text{SE}}(\hat{S}(t)) \approx \hat{S}(t) \sqrt{\sum_{j=1}^i \frac{m_j}{n_j(n_j - m_j)}}$$

for $t_{(i)} \leq t < t_{(i+1)}$

Greenwood SE/CI for KM

- For the leukemia example,

$$\widehat{\text{SE}}(\hat{S}(6)) = 0.8571 \sqrt{\frac{3}{21 \cdot (21 - 3)}} = 0.0764$$

$$\widehat{\text{SE}}(\hat{S}(7)) = 0.8067 \sqrt{\frac{3}{21 \cdot 18} + \frac{1}{17 \cdot 16}} = 0.0869$$

- An approximate $100(1 - \alpha)\%$ CI is

$$\hat{S}(t) \pm z_{1-\alpha/2} \widehat{\text{SE}}(\hat{S}(t))$$

Greenwood SE/CI for KM

- Greenwood based CIs are symmetric
- This is problematic when the survival function is near 0 or 1 because it is possible for part of the CI to lie outside the interval $[0,1]$
- Pragmatic solution: Set relevant end of interval to 0 or 1 in this case
- Many other methods exist to estimate the standard error and obtain confidence intervals
- All have pointwise interpretation; different methods exist to obtain *confidence bands*

Testing

- How do we test under minimal assumptions whether two survival functions are different?
- For example: Suppose leukemia patients are randomized to treatment or placebo. Are the survival functions the same between the two groups?
- Without censoring, use a rank test (e.g., Wilcoxon rank sum test)
- In the presence of right censoring, use the log-rank test

Log-Rank Test

- Suppose we have data from two samples

$$(T_{ij}, \delta_{ij})$$

for $i = 1, 2$ and $j = 1, 2, \dots, n_i$

- We want to test

$$H_0 : S_1(t) = S_2(t) \text{ for all } t$$

where

$$S_j(t) = \Pr[T_j^* > t] \text{ for } j = 1, 2$$

Log-Rank Test

- Let $t_{(1)}, t_{(2)}, \dots, t_{(K)}$ be the distinct ordered failure times in the two groups combined
- At each time $t_{(k)}$, construct the table:

Group	At risk	Events	Survive
1	$R_1(t_{(k)})$	m_{1k}	$R_1(t_{(k)}) - m_{1k}$
2	$R_2(t_{(k)})$	m_{2k}	$R_2(t_{(k)}) - m_{2k}$
	$R(t_{(k)})$	m_k	$R(t_{(k)}) - m_k$

Log-Rank Test

- Under H_0 , the expected number of deaths in group 1 is

$$E_{1k} = R_1(t_{(k)}) \frac{m_k}{R(t_{(k)})}$$

- The hypergeometric variance is

$$V_{1k} = \frac{R_1(t_{(k)})R_2(t_{(k)})m_k(R(t_{(k)}) - m_k)}{R(t_{(k)})^2(R(t_{(k)}) - 1)}$$

Log-Rank Test

- The log-rank (Mantel-Haenszel) statistic uses

$$E_1 = \sum_{k=1}^K E_{1k}, \quad O_1 = \sum_{k=1}^K m_{1k}, \quad V_1 = \sum_{k=1}^K V_{1k}$$

- Under $H_0 : S_1(t) = S_2(t)$ for all t ,

$$X = \frac{(O_1 - E_1)^2}{V_1} \sim \chi_1^2$$

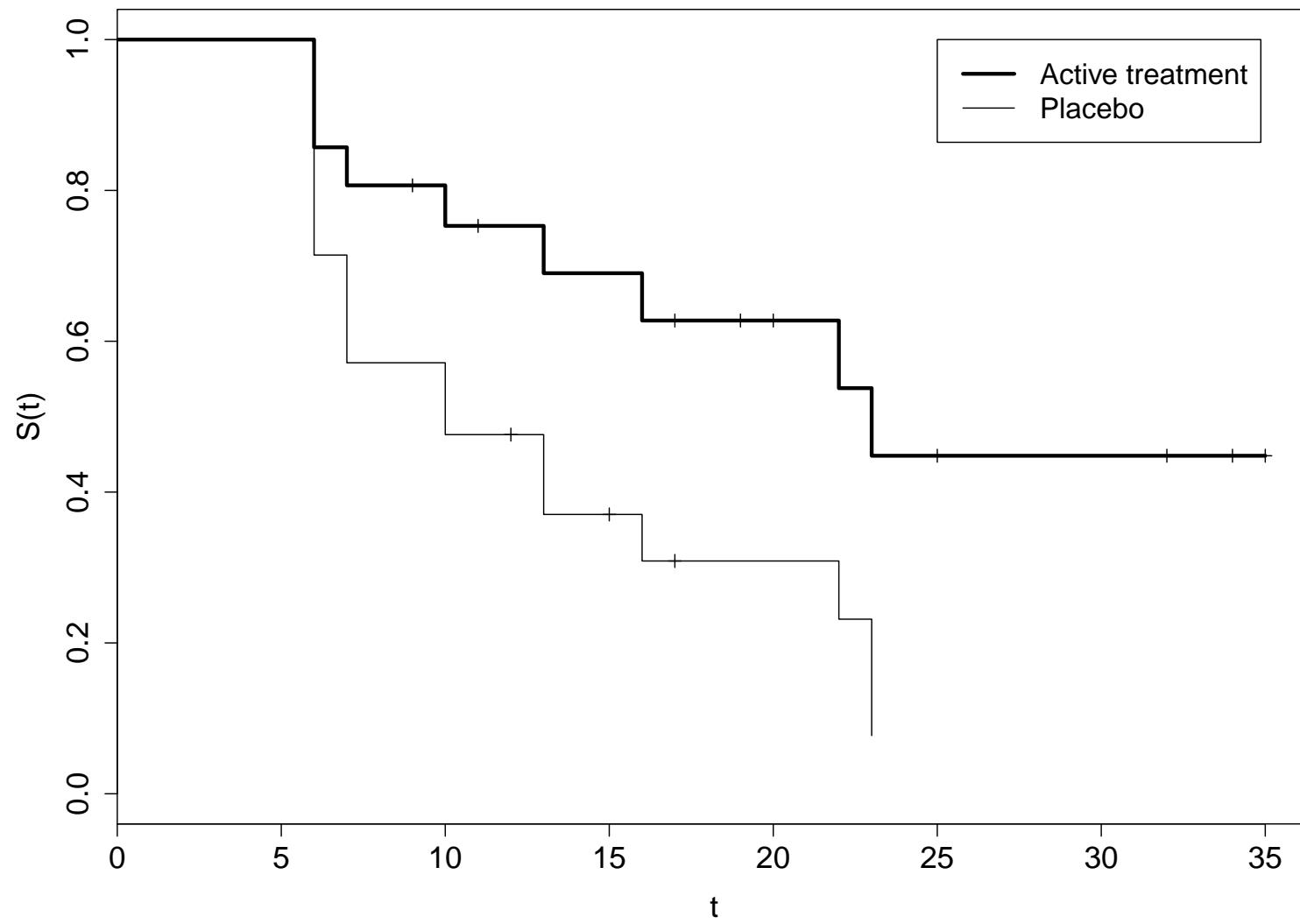
Log-Rank Test

- Leukemia example

Treatment ($n = 21$)	Placebo ($n = 21$)
6, 6, 6, 6+	6, 6, 6, 6
7, 9+, 10, 10+	6, 6, 7, 7
11+, 13, 16, 17+	7, 10, 10, 12+, 13
19+, 20+, 22, 23	13, 15+, 16, 17+
25+, 32+, 32+, 34+, 35+	22, 23, 23, 23+

where + indicates that the person was censored at that time

Log-Rank Test: Leukemia Example



Code for Plotting Kaplan-Meier Curves

- R

```
library("survival")
fit <- survfit(Surv(t, delta)~rx,conf.type="none")
pdf("surv_leuk1.pdf",width=11,height=8.5)
plot(fit,xlab="t",ylab="S(t)",lwd=c(1,3))
legend(25,1,c("Active treatment","Placebo"),lwd=c(3,1))
dev.off()
```

- SAS

```
proc lifetest plots=(s) graphics;
  time t*delta(0);
  strata trt;
```

Log-Rank Test “By Hand”: Leukemia Example

$t_{(k)}$	m_{1k}	$R_1(t_{(k)})$	m_{2k}	$R_2(t_{(k)})$	m_k	$R(t_{(k)})$	E_{1k}	V_{1k}
6	3	21	6	21	9	42	4.50	1.81
7	1	17	3	15	4	32	2.13	0.90
10	1	15	2	12	3	27	1.67	0.68
13	1	12	2	9	3	21	1.71	0.66
16	1	11	1	6	2	17	1.29	0.43
22	1	7	1	4	2	11	1.27	0.42
23	1	6	2	3	3	9	2.00	0.50
						14.57	5.4	
						9		

Log-Rank Test: Leukemia Example

- Therefore

$$X = \frac{(9 - 14.57)^2}{5.4} = 5.75$$

$$\Pr[\chi_1^2 > 5.75] = 0.0165$$

- R code:

```
> survdiff(Surv(t, delta)~rx)
```

Call:

```
survdiff(formula = Surv(t, delta) ~ rx)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
rx=p	21	17	11.4	2.72	5.75
rx=t	21	9	14.6	2.13	5.75

Chisq= 5.8 on 1 degrees of freedom, p= 0.0165

Log-Rank Test: Leukemia Example cont.

- SAS code

```
proc lifetest;  
  time t*delta(0);  
  strata trt;
```

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	5.7507	1	0.0165
Wilcoxon	4.3357	1	0.0373
-2Log(LR)	6.0441	1	0.0140

Log-Rank Test: SAS

- We can also use proc freq and the Mantel-Haenszel statistic, setting up a 2×2 table at each time point at which there is at least one event. All those in the risk set at such a time contribute to the table at that time

```
data;  
    input time group remission wt;  
cards;  
6 1 1 3  
6 1 0 18  
6 2 1 6  
6 2 0 15  
7 1 1 1  
7 1 0 16  
7 2 1 3  
7 2 0 12  
.  
.  
.
```

Log-Rank Test: SAS cont.

```
proc freq order=data;  
  tables time*group*remission / chisq cmh;  
  weight wt;
```

The FREQ Procedure

Summary Statistics for group by remission
Controlling for time

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob

1	Nonzero Correlation	1	5.7507	0.0165
2	Row Mean Scores Differ	1	5.7507	0.0165
3	General Association	1	5.7507	0.0165

Cox / Proportional Hazards Model

- The *hazard function* $\lambda(t)$ is the instantaneous event rate at any time t

$$\lambda(t) = \lim_{\Delta t \rightarrow 0^+} \frac{Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)}$$

- The proportional hazards model is a linear model for the log of the hazard or, equivalently, a multiplicative model for the hazard

$$\log \lambda(t) = \log \lambda_0(t) + \beta X$$

or

$$\lambda(t) = \lambda_0(t) \exp(\beta X)$$

- $\lambda_0(t)$ is called the *baseline hazard*

Cox / Proportional Hazards Model

- Consider two values of X , x_1 and x_2 ; then

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t) \exp(\beta x_1)}{\lambda_0(t) \exp(\beta x_2)} = \frac{e^{\beta x_1}}{e^{\beta x_2}}$$

independent of t

- This independence of t is an *assumption* and needs to be checked
- Let X be an indicator of being in one of two exposure or treatment groups, then if $x_1 = 1$ and $x_2 = 0$,

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{e^{\beta \cdot 1}}{e^{\beta \cdot 0}} = e^\beta$$

- e^β is the *hazard ratio* comparing group 1 to group 2

Leukemia Treatment Example: R

- There are a substantial number of tied observations;
R and SAS have different defaults for handling ties
- Using R's default method of handling ties (Efron)

```
> summary(coxph(Surv(t, delta)~rx))
Call:
coxph(formula = Surv(t, delta) ~ rx)

n= 42

      coef  exp(coef)  se(coef)      z Pr(>|z|)
rxt -0.9684    0.3797   0.4164 -2.325   0.0200 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

      exp(coef)  exp(-coef) lower .95 upper .95
rxt     0.3797      2.634    0.1679     0.8588
```

Leukemia Treatment Example: SAS

- Using the “exact” method for ties rather than the SAS default (Breslow)

```
proc phreg;  
  model t*delta(0) = active / ties=exact;
```

Summary of the Number of Event and Censored Values

	Total	Event	Censored	Percent Censored
	42	26	16	38.10

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
active	1	-0.97790	0.41896	5.4482	0.0196	0.376

BIOS 662 Fall 2018

Survey Sampling, Part I

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Preliminaries
- Simple random sampling
 - Population mean
 - Population total
 - Sample size
 - Proportion

Preliminaries: References

- SK Thompson. *Sampling*. John Wiley and Sons, 1992
- L Kish. *Survey sampling*. Wiley, New York, 1965
- WG Cochran. *Sampling Techniques*. John Wiley and Sons, 1977

Preliminaries: What is “(Survey) Sampling”?

- *Sampling* study: Selecting some part of a population to be observed so that one may estimate something about the whole of the population
- Example: To estimate the amount of lichen in a well-defined area, a biologist collects lichen from selected small plots within the study area
- Typically want to estimate total or mean
- Observational – does not intentionally perturb or disturb population (i.e., not experimental)
- However, one does have control over how the sample is selected

Preliminaries: Terminology

- *Population*: The group of units (e.g., people) we are sampling and studying; assumed to be of known, finite size, rather than infinite
- *Sampling design*: The strategy followed in selecting a sample from a population
- *Sampling unit*: Unit designated for listing and selection in a sample survey (e.g., persons, dwellings, households, area units, pharmacies)
- *Sampling frame*: List of sampling units from which a sample is drawn

Preliminaries: Terminology

- *Variable*: Some measurement taken on members of the sample (e.g., number of children ever born to a woman aged 15-49 years); sometimes call this the y -variable or x -variable
- *Selection probability*: Likelihood, over repeated applications of a sampling design, that a particular unit will be chosen for a sample
- *Probability sampling*
 - Sampling in which the design calls for using random methods to ultimately decide which units are chosen
 - Every unit has a known, non-zero selection probability

Preliminaries: Terminology

- *Equal-probability sampling*
 - Probability sampling in which all units in the population have the same selection probability
 - Also known as “self-weighted” sampling or “epsem” (equal probability of selection method) sampling
- *Non-probability sampling*
 - Sampling in which subjective judgment (usually by interviewers) is used to decide who is chosen in the sample
 - Selection probabilities cannot be determined
 - Difficult to determine if the sample is representative

Preliminaries: Terminology

- *Unbiased* estimator: An estimator which, if repeated over all possible samples that might be selected using the sampling design, would yield estimates which on average equal the parameter being estimated (e.g., sample mean from a simple random sample is an unbiased estimator of the population mean)
- Also known as *design-unbiased*
- Key idea: the randomness in the estimator is induced by the sampling design

Preliminaries: Software

- SAS: proc surveymeans, surveyfreq, ...
- R: “survey” package

Preliminaries: Sampling Designs

- Simple random sampling
- Stratified sampling
- Cluster sampling

Simple Random Sampling (SRS)

- Let N denote the number of units in the population
- *Simple random sampling*, or random sampling *without replacement* (SRSWOR), is the sampling design in which n distinct units are selected from the N units in the population in such a way that every possible combination of the n units is equally likely to be the sample selected
- A simple random sample can be obtained through a sequence of independent selections from the whole population such that each unit has an equal probability of selection at each step, discarding repeat selections and continuing until n distinct units are obtained
- $f \equiv n/N$ is the sampling rate or sampling fraction

Obtaining a Simple Random Sample

- A. Number the units in the population (i.e., sampling frame) from 1 to N .
- B. Select and record a random number between 1 and N .
- C. At each subsequent step, select a random integer between 1 and N . If it is the same as a previously selected number, discard it. Otherwise, record it.
- D. Continue in this manner until n different numbers between 1 and N have been chosen.
- E. Population units corresponding to the selected numbers form a simple random sample of size n .

Obtaining a Simple Random Sample

Alternative approach:

- A. Generate a random number from $U(0, 1)$ for each unit in the population (i.e., sampling frame).
- B. Sort in order of the random numbers.
- C. Take the first n units in the sorted list.

Key Properties of SRS

- All possible simple random samples have the same chance of being selected
- The probability that any one population unit will be chosen is n/N
- Selection probabilities in an SRS are not statistically independent

$$\Pr[i \text{ in sample}] = \frac{\binom{1}{1} \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{\frac{(N-1)!}{(n-1)!(N-n)!}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N}$$

$$\Pr[i \text{ and } j \text{ in sample}] = \frac{\binom{2}{2} \binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)} \neq \left(\frac{n}{N}\right)^2$$

SRS: Estimating Population Mean

- Denote the (finite) population mean by

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i$$

- Denote the (finite) population variance by

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2$$

- Let Z_i indicate whether unit i is in the sample,
that is $Z_i = 1$ if i is sampled, $Z_i = 0$ otherwise
- Key point: The y_i are fixed, the Z_i are random

SRS: Estimating Population Mean

- Sample mean

$$\bar{y} = \frac{1}{n} \sum_{i=1}^N y_i Z_i$$

- Sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^N (y_i - \bar{y})^2 Z_i$$

- The sample mean is unbiased: Each Z_i is Bernoulli with $E(Z_i) = n/N$, thus

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^N y_i E(Z_i) = \frac{1}{N} \sum_{i=1}^N y_i = \mu$$

SRS: Estimating Population Mean

- To derive the variance of the sample mean,

$$\text{Var}(\bar{y}) = \frac{1}{n^2} \left(\sum_{i=1}^N y_i^2 \text{Var}(Z_i) + \sum_{i \neq j} y_i y_j \text{Cov}(Z_i, Z_j) \right)$$

we need the variance and covariance terms

- The variance is easy because the Z_i are Bernoulli

$$\text{Var}(Z_i) = \frac{n}{N} \left(1 - \frac{n}{N} \right)$$

- For SRS, the Z_i are not independent

$$\begin{aligned} \text{Cov}(Z_i, Z_j) &= E(Z_i Z_j) - E(Z_i)E(Z_j) = \frac{n}{N} \frac{(n-1)}{(N-1)} - \left(\frac{n}{N} \right)^2 \\ &= -\frac{n}{N} \left(1 - \frac{n}{N} \right) \frac{1}{N-1} \end{aligned}$$

SRS: Estimating Population Mean

- Thus

$$\text{Var}(\bar{y}) = \frac{1}{n^2} \left(\frac{n}{N} \right) \left(1 - \frac{n}{N} \right) \left(\sum_{i=1}^N y_i^2 - \frac{1}{N-1} \sum_{i \neq j} y_i y_j \right)$$

- Using the identity

$$\sum_{i=1}^N (y_i - \mu)^2 = \sum_{i=1}^N y_i^2 - \frac{(\sum y_i)^2}{N} = \frac{1}{N} \left((N-1) \sum_{i=1}^N y_i^2 - \sum_{i \neq j} y_i y_j \right)$$

we get

$$\text{Var}(\bar{y}) = \frac{1}{n} \left(1 - \frac{n}{N} \right) \frac{\sum (y_i - \mu)^2}{N-1} = \left(1 - \frac{n}{N} \right) \frac{\sigma^2}{n}$$

SRS: Estimating Population Mean

- The quantity

$$1 - \frac{n}{N} = \frac{N - n}{N} = 1 - f$$

is called the *finite population correction factor*

- If the population is large relative to the sample size, n/N will be small, so that

$$\text{Var}(\bar{y}) \approx \frac{\sigma^2}{n}$$

- On the other hand, $\text{Var}(\bar{y}) \rightarrow 0$ as $n \rightarrow N$

SRS: Estimating Population Variance

- Exercise: Show that $E(s^2) = \sigma^2$, i.e., the sample variance, is an unbiased estimator for the finite population variance
- From this fact, it follows that an unbiased estimator for $\text{Var}(\bar{y})$ is given by

$$\widehat{\text{Var}}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n}$$

SRS: Estimating Population Total

- Define the population total

$$\tau = \sum_{i=1}^N y_i = N\mu$$

- Unbiased estimator

$$\hat{\tau} = N\bar{y} = \frac{N}{n} \sum_{i=1}^N y_i Z_i$$

with variance

$$\text{Var}(\hat{\tau}) = N^2 \text{Var}(\bar{y}) = N(N-n) \frac{\sigma^2}{n}$$

- Unbiased estimator of variance

$$\widehat{\text{Var}}(\hat{\tau}) = N^2 \widehat{\text{Var}}(\bar{y}) = N(N-n) \frac{s^2}{n}$$

SRS: Estimating Population Total

- The estimator is often written as

$$\hat{\tau} = \sum_{i=1}^N w_i y_i Z_i = \sum_{i=1}^N \frac{y_i Z_i}{\pi_i}$$

where $w_i^{-1} = \pi_i = n/N = f$ is the selection probability

- “Inverse probability weighting”
- This is the formulation SAS uses (more below)
- Special case of the *Horvitz-Thompson* estimator

(Horvitz, D.G. and Thompson, D.J. (1952) A generalization of sampling without replacement from a finite universe. J. Amer. Stastist. Assoc., 47, 663-685.)

SRS: Finite-population CLT

- The usual central limit theorem requires independence
- Imagine a sequence of populations with population size N becoming large along with sample size n . Let μ_N be the population mean and \bar{y}_N the sample mean for an SRS from that population.
- According to the finite-population CLT

$$\frac{\bar{y}_N - \mu_N}{\sqrt{\text{Var}(\bar{y}_N)}} \rightarrow Z \sim N(0, 1)$$

as both $n \rightarrow \infty$ and $N - n \rightarrow \infty$

SRS: Finite-population CIs

- This leads to approximate $100(1 - \alpha)\%$ CIs for the population mean μ

$$\bar{y} \pm t_{n-1,1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}$$

- Likewise, for the population total τ

$$\hat{\tau} \pm t_{n-1,1-\alpha/2} \sqrt{N(N-n) \frac{s^2}{n}}$$

SRS: Example

- (Section 2.3, Thompson 1992)
- A survey of the caribou population was done by aircraft
- A 286 mile wide study region was divided into exhaustive and mutually exclusive one-mile strips ($N = 286$)
- An SRS of $n = 15$ yielded counts of 1, 2, 4, 4, 5, 7, 10, 15, 21, 21, 29, 36, 50, 86, 98
- Sample mean $\bar{y} = 25.933$
- Sample variance $s^2 = 919.067$

SRS: Example cont.

- Thus

$$\widehat{\text{Var}}(\bar{y}) = \left(1 - \frac{15}{286}\right) \frac{919.067}{15} = 58.058$$

yielding 95% CI for μ

$$25.933 \pm 2.145\sqrt{58.058} = (9.59, 42.28)$$

- For the population total, $\hat{\tau} = N\bar{y} = 7417$, with 95% CI

$$7417 \pm 2.145\sqrt{286(286 - 15)\frac{919.067}{15}} = (2743, 12091)$$

SRS: Example Using SAS

```
proc means mean clm sum; *** Assuming infinite population;  
var counts;
```

	Lower 95%	Upper 95%	
Mean	CL for Mean	CL for Mean	Sum
25.9333333	9.1448299	42.7218368	389.0000000

```
proc surveymeans total=286 mean clm sum clsum;  
var counts;
```

	Std Error		
Variable	Mean	of Mean	95% CL for Mean
counts	25.933333	7.619553	9.59101702 42.2756497

Variable	Sum	Std Dev	95% CL for Sum
counts	389.000000	114.293298	143.865255 634.134745

SRS: Example Using SAS

```
proc surveymeans total=286 mean clm sum clsum;  
  var counts;  
  weight wt; *** wt=286/15 for all;
```

Variable	Mean	Std Error of Mean	95% CL for Mean
counts	25.933333	7.619553	9.59101702 42.2756497

Variable	Sum	Std Dev	95% CL for Sum
counts	7416.933333	2179.192221	2743.03087 12090.8358

SRS: Example Using R

```
> library("survey")

> caribou <- data.frame(y=c(1,2,4,4,5,7,10,15,21,21,29,36,50,86,98),fpc=15/286)

> design <- svydesign(ids=~1,data=caribou,fpc=~fpc)

> # R uses Z not t for the confidence intervals here
> svymean(caribou$y, design); confint(svymean(caribou$y, design))

      mean      SE
y  25.933333 7.6196

      2.5 %      97.5 %
y  10.99928344 40.86738323

> svytot(caribou$y, design); confint(svytot(caribou$y, design))

      total      SE
y    7417 2179.2

      2.5 %      97.5 %
y  3145.795 11688.07
```

SRS: Sample Size

- Suppose we want to choose the smallest sample size n such that

$$\Pr[|\hat{\theta} - \theta| > d] \leq \alpha$$

- Assuming

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim N(0, 1)$$

choose n such that

$$z_{1-\alpha/2} \sqrt{\text{Var}(\hat{\theta})} = d$$

SRS: Sample Size

- For example, if $\theta = \mu$, choose n such that

$$z_{1-\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}} = d$$

- This implies

$$n = \frac{1}{1/n_0 + 1/N} \quad \text{where} \quad n_0 = \frac{z_{1-\alpha/2}^2 \sigma^2}{d^2}$$

- Note that if $N \gg n$, then $n \approx n_0$

SRS: Sample Size

- If $\theta = \tau$, choose n such that

$$z_{1-\alpha/2} \sqrt{N(N-n) \frac{\sigma^2}{n}} = d$$

implying

$$n = \frac{1}{1/n_0 + 1/N} \quad \text{where} \quad n_0 = \frac{N^2 z_{1-\alpha/2}^2 \sigma^2}{d^2}$$

- Example: Find n necessary to estimate the caribou population total to within 2,000 animals of the true total with 90% confidence

Here $\sigma^2 = 919$, $d = 2000$, $\alpha = 0.1$

So $n_0 = 50.9$, $n = 43.2$

SRS: Estimating a Proportion

- Suppose responses are binary, e.g., want to estimate the proportion of voters favoring a candidate for elected office
- Let $y_i = 1$ if unit i has the attribute of interest,
 $y_i = 0$ otherwise
- Then μ is the proportion of units in the population with the attribute
- Thus can use methods from before. However, there are some special features now:
 - Formulas simplify considerably
 - Exact confidence intervals are possible
 - Sample size calculation does not require information about population parameters

SRS: Estimating a Proportion

- Let the proportion of the population with the attribute be

$$p = \frac{1}{N} \sum_{i=1}^N y_i = \mu$$

- Finite population variance

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (y_i - p)^2}{N - 1} = \frac{\sum_{i=1}^N y_i^2 - Np^2}{N - 1} \\ &= \frac{Np - Np^2}{N - 1} \\ &= \frac{Np(1 - p)}{N - 1}\end{aligned}$$

SRS: Estimating a Proportion

- Proportion in the sample with the attribute

$$\hat{p} = \frac{1}{n} \sum_{i=1}^N y_i Z_i = \bar{y}$$

- Sample variance

$$\begin{aligned}s^2 &= \frac{\sum_{i=1}^N (y_i - \bar{y})^2 Z_i}{n - 1} \\&= \frac{\sum_{i=1}^N y_i^2 Z_i - n\hat{p}^2}{n - 1} \\&= \frac{n\hat{p}(1 - \hat{p})}{n - 1}\end{aligned}$$

SRS: Estimating a Proportion

- Because the sample proportion is a sample mean of an SRS, all previous results hold; in particular:

$$E(\hat{p}) = p$$

$$\text{Var}(\hat{p}) = \left(\frac{N-n}{N-1} \right) \frac{p(1-p)}{n}$$

$$\widehat{\text{Var}}(\hat{p}) = \left(\frac{N-n}{N} \right) \frac{\hat{p}(1-\hat{p})}{n-1}$$

$$E(\widehat{\text{Var}}(\hat{p})) = \text{Var}(\hat{p})$$

SRS: CI for a Proportion

- Approximate $100(1 - \alpha)\%$ CI

$$\hat{p} \pm t_{1-\alpha/2, n-1} \sqrt{\widehat{\text{Var}}(\hat{p})}$$

- The approximation improves as n increases and the closer p is to 0.5
- An exact CI can be computed based on inverting a test and the hypergeometric distribution

SRS: Exact CI for a Proportion

- Suppose ν units in the population have the attribute of interest
- Let

$$X = \sum_{i=1}^N y_i Z_i$$

- Then

$$\Pr[X = j | \nu] = \frac{\binom{\nu}{j} \binom{N-\nu}{n-j}}{\binom{N}{n}}$$

SRS: Exact CI for a Proportion

- Suppose we observe $X = x$ for one particular SRS (such that $\hat{p} = x/n$)
- Let ν_L be the smallest integer such that

$$\Pr[X \geq x | \nu_L] > \alpha/2$$

and let ν_U be the largest integer such that

$$\Pr[X \leq x | \nu_U] > \alpha/2$$

- Then an exact $100(1 - \alpha)\%$ CI is given by

$$(\nu_L/N, \nu_U/N)$$

SRS: Sample Size for a Proportion

- To obtain an estimator \hat{p} having probability at least $1 - \alpha$ of being no farther than d from the population proportion

$$n = \frac{Np(1-p)}{(N-1)d^2/z_{1-\alpha/2}^2 + p(1-p)}$$

- If $N \gg n$

$$n \approx \frac{z_{1-\alpha/2}^2 p(1-p)}{d^2}$$

- If no a-priori knowledge of p , conservatively assume $p = 0.5$

SRS: Estimating a Ratio

- Example 1: A biologist studying an animal population selects an SRS of plots in the study region. In each selected plot, she counts the number y_i of young animals and the number x_i of adult females, with the object of estimating the ratio of young to adult females in the population
- Example 2: In a household survey to estimate the number of television sets per person in the region, an SRS of households is conducted. For each selected household the number y_i of television sets and the number x_i of people is recorded

SRS: Estimating a Ratio

- Ratio estimator

$$r = \frac{\sum_{i=1}^N y_i Z_i}{\sum_{i=1}^N x_i Z_i} = \frac{\bar{y}}{\bar{x}}$$

- Note that the denominator of the estimator is a random variable

SRS: Concluding Remarks

- SRS is the simplest probability sampling method
- It is quite rarely used in practice
- Exception: When the sample size and population size are small and stratified sampling is not possible

BIOS 662 Fall 2018

Survey Sampling, Part II

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- Stratified sampling
 - Introduction
 - Notation and estimands
 - Estimators
 - Allocation strategies
 - Example

Stratified Sampling

- *Stratification:* The process of dividing a population of units into distinct sub-populations called strata
- Strata are formed so that each population unit is assigned to only one stratum
- To draw a sample of US counties, we might stratify by region (NE, SE, NW, SW, ...)
- How is stratification used in sample surveys?

Stratified Sampling

- The population is divided into H strata so that each population unit is a member of only one stratum.
- Let N_h denote the number of population units in stratum h for $h = 1, \dots, H$.
- Thus the total number of units in the population is

$$N = \sum_{h=1}^H N_h$$

- Let n_h denote the sample size for stratum h , so that the total sample size is

$$n = \sum_{h=1}^H n_h$$

Stratified Sampling

- A sample of size n_h is selected by some probability design (e.g., SRS) from each of the H strata, independent of each other
- Stratum-specific parameters (e.g., means, totals) are estimated separately using data from each of the H strata
- An estimate of the population parameter is produced by appropriately combining the H individual stratum estimates
- If SRS is used within stratum, this is called *stratified random sampling*

Notation and Estimands

- Let y_{hi} denote the variable of interest associated with unit i of stratum h ($i = 1, \dots, N_h$; $h = 1, \dots, H$)
- Let $Z_{hi} = 1$ if the corresponding unit is in the sample, 0 otherwise
- Stratum total

$$\tau_h = \sum_{i=1}^{N_h} y_{hi}$$

- Population total

$$\tau = \sum_{h=1}^H \tau_h = \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi}$$

Notation and Estimands

- Stratum mean

$$\mu_h = \frac{\tau_h}{N_h} = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$$

- Population mean

$$\mu = \frac{\tau}{N} = \frac{\sum_h \sum_i y_{hi}}{N} = \sum_h \frac{N_h}{N} \cdot \frac{1}{N_h} \sum_i y_{hi} = \sum_h W_h \mu_h$$

where $W_h = N_h/N$ is the proportion of population units in stratum h

Population Mean Estimator

- Estimator of the population mean

$$\bar{y} = \sum_h W_h \bar{y}_h$$

where \bar{y}_h is an estimator of the mean μ_h for stratum h

- $E(\bar{y}_h) = \mu_h$ implies $E(\bar{y}) = \mu$

- Estimator of the variance of \bar{y}

$$\widehat{\text{Var}}(\bar{y}) = \sum_h W_h^2 \widehat{\text{Var}}(\bar{y}_h)$$

- $E(\widehat{\text{Var}}(\bar{y}_h)) = \text{Var}(\bar{y}_h)$ implies $E(\widehat{\text{Var}}(\bar{y})) = \text{Var}(\bar{y})$

Population Mean Estimator

- If stratified random sampling, then

$$\bar{y}_h = \frac{\sum_i y_{hi} Z_{hi}}{n_h}$$

and

$$\widehat{\text{Var}}(\bar{y}) = \sum_h W_h^2 \left(\frac{1 - f_h}{n_h} \right) s_h^2$$

where $f_h = n_h/N_h$ is the stratum-specific sampling rate
and s_h^2 is the within-stratum sample variance

Population Mean Estimator

- CIs

$$\bar{y} \pm t_{1-\alpha/2, df} \sqrt{\widehat{\text{Var}}(\bar{y})}$$

where

$$df = \frac{(\sum_h a_h s_h^2)^2}{\sum_h (a_h s_h^2)^2 / (n_h - 1)}$$

and

$$a_h = N_h(N_h - n_h)/n_h$$

- If all the N_h are equal and all the n_h are equal, then

$$df = n - H$$

Population Total Estimator

- Estimator of population total

$$\hat{\tau} = N\bar{y} = \sum_h N_h \bar{y}_h$$

- $E(\bar{y}_h) = \mu_h$ implies $E(\hat{\tau}) = \tau$

- Estimator of variance

$$\widehat{\text{Var}}(\hat{\tau}) = N^2 \widehat{\text{Var}}(\bar{y}) = \sum_h N_h^2 \left(\frac{1 - f_h}{n_h} \right) s_h^2$$

with the second equality assuming stratified random sampling

Population Total Estimator

- $E(\widehat{\text{Var}}(\bar{y}_h)) = \text{Var}(\bar{y}_h)$ implies $E(\widehat{\text{Var}}(\hat{\tau})) = \text{Var}(\hat{\tau})$

- CIs

$$\hat{\tau} \pm t_{1-\alpha/2, df} \sqrt{\widehat{\text{Var}}(\hat{\tau})}$$

where df is as specified on the previous page

Population Total Proportion

- Estimator of population proportion

$$\hat{p} = \sum_h W_h \hat{p}_h$$

where the \hat{p}_h are the stratum-specific estimators;

- \hat{p} is a special case of \bar{y}
- Estimator of variance for stratified random sampling

$$\widehat{\text{Var}}(\hat{p}) = \sum_h W_h^2 \left(\frac{1 - f_h}{n_h - 1} \right) \hat{p}_h (1 - \hat{p}_h)$$

Stratification Principle

- Variances depend on within-stratum population variance terms only
- Thus estimators will be more precise the smaller

$$\sigma_h^2 = \sum_i (y_{hi} - \mu_h)^2 / (N_h - 1)$$

- That is, estimation of the population mean or total will be most precise if the population is partitioned into strata in such a way that *within each stratum, the units are as similar as possible*
- For example, in a survey of a plant or animal population, the study area might be stratified into regions of similar habitat or elevation, because we expect abundancies to be more similar within strata than between strata

Stratification Principle: Example

- Suppose $N = 6$; $H = 2$; $N_h = 3$ for $h = 1, 2$
- Stratum 1 values: 0, 1, 2; stratum 2 values: 4, 5, 9
- Population variance $\sigma^2 = 10.7$
- Stratum variances $\sigma_1^2 = 1$, $\sigma_2^2 = 7$
- For SRS with $n = 4$,

$$\text{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n} = \left(1 - \frac{4}{6}\right) \frac{10.7}{4} = 0.89$$

- For stratified random sampling with $n_1 = n_2 = 2$,

$$\text{Var}(\bar{y}) = \left(\frac{3}{6}\right)^2 \left(\frac{1 - 2/3}{2}\right) 1 + \left(\frac{3}{6}\right)^2 \left(\frac{1 - 2/3}{2}\right) 7 = 0.33$$

Allocation Strategies

- How to choose the sample size n_h for each stratum?
- Four strategies
 - Proportionate: same sampling rates
 - Optimum: most cost efficient
 - Balanced: equal sample sizes
 - Disproportionate: unequal sampling rates (to oversample important domains)

Proportionate Stratified Sampling

- Same sampling rate f_h for all strata:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

- Equivalently

$$W_h = \frac{N_h}{N} = \frac{n_h}{n} = w_h$$

- The proportion of the sample chosen from any given stratum will be the same as the proportion of the population in that stratum

Proportionate Stratified Sampling

- Each unit in the population has the same probability of selection
- This type of design is called a *self-weighting design* because sample estimates of population mean and proportion are simple arithmetic means
- For example, the population mean estimator for proportionate stratified random sampling is

$$\bar{y} = \frac{1}{n} \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi} Z_{hi}$$

Stratification Principle

- Claim: The variance of estimators from proportionate stratified random sampling are always less than or equal to the variance of estimators from SRS
- Sketch of proof (see Cochran 1977, pages 99-100)

$$\begin{aligned}(N - 1)\sigma^2 &= \sum_h \sum_i (y_{hi} - \mu)^2 \\&= \sum_h \sum_i (y_{hi} - \mu_h)^2 + \sum_h N_h (\mu_h - \mu)^2 \\&= \sum_h (N_h - 1)\sigma_h^2 + \sum_h N_h (\mu_h - \mu)^2\end{aligned}$$

implying

$$\sigma^2 \approx \sum_h W_h \sigma_h^2 + \sum_h W_h (\mu_h - \mu)^2$$

Stratification Principle

- Thus

$$\text{Var}_{\text{SRS}}(\bar{y}) = (1 - f) \sigma^2 / n$$

$$\begin{aligned} &\approx (1 - f) \sum_h W_h \sigma_h^2 / n \\ &\quad + (1 - f) \sum_h W_h (\mu_h - \mu)^2 / n \end{aligned}$$

- Under proportionate stratified random sampling

$$\text{Var}_{\text{pro}}(\bar{y}) = (1 - f) \sum_h W_h^2 \sigma_h^2 / n_h = (1 - f) \sum_h W_h \sigma_h^2 / n$$

- Therefore

$$\text{Var}_{\text{SRS}}(\bar{y}) \approx \text{Var}_{\text{pro}}(\bar{y}) + (1 - f) \sum_h W_h (\mu_h - \mu)^2 / n$$

Stratification Principle

- For this reason, proportionate stratified random sampling is often considered the default
- Gains over SRS will be greatest if strata are internally homogeneous

General Guidelines

- The stratification variable should be highly correlated with the principal characteristic being measured in the survey (e.g., age may be a good stratification variable if we are doing a survey on limitation due to chronic illness)
- Strata should be internally homogeneous
- The variance of a population estimate will be smallest (for fixed cost) when each stratum sampling rate is directly related to the variability of units within the stratum and inversely related to the unit cost of data collection in the stratum
- Proportionate stratified sampling is always “safe” in that precision will never be worse than for SRS

General Advantages

- Improved precision of estimates (i.e., smaller variances), which leads to narrower confidence intervals
- Better control of sample sizes for sub-populations which can be defined by strata and for which separate estimates may be sought
- Sampling designs can be made more flexible
- For example, special strata may be established to handle segments of the population that are more difficult to survey (e.g., transient populations in household surveys)

Note of Caution

- Several stratum allocations yield very close to the optimum allocation
- Excessive attempts to determine the actual optimum allocation is almost never cost-effective.

Example of Analysis

- A county with two relatively large communities wants to do a survey on certification of the emergency medical technicians (EMTs) who work in the county and who are required to take special training and pass a competency exam for periodic certification
- Most EMTs work in “City A,” which is relatively large and is located in the main urban area of the county; “City B” is smaller and has fewer EMTs; and the rest of the county’s EMTs work in smaller towns and in rural areas
- Because of suspected similarities in certification patterns among EMTs in City A and comparable similarities in City B, we decide to divide the county into three strata, “City A”, “City B” and “Other”

Example

- We want to estimate:
 - μ , the average number of hours of certification training in the year prior to the last certification
 - μ_1 , the average number of hours of certification training in City A in the year prior to the last certification
 - τ , the total number of certification hours for EMTs in the county for the year prior to the last certification.
 - p , the proportion of EMTs who passed their last periodic certification exam on the first attempt

Example

- Use proportionately allocated sample sizes

h	Stratum Composition	N_h	n_h
1	City A	155	20
2	City B	62	8
3	Rural Area	93	12
Total		310	40

Example Data

Stratum 1			Stratum 2			Stratum 3		
i	Hours	Passed	i	Hours	Passed	i	Hours	Passed
1	35	1	11	29	0	1	27	1
2	28	1	12	31	1	2	4	0
3	26	1	13	39	1	3	49	0
4	41	1	14	38	0	4	10	1
5	43	1	15	40	0	5	15	0
6	29	0	16	45	1	6	41	0
7	32	1	17	28	1	7	25	0
8	37	1	18	27	1	8	30	0
9	36	1	19	35	1			9 12 1
10	25	1	20	34	1			10 32 0
								11 34 0
								12 24 1

Example

- Summary statistics

Stratum 1	Stratum 2	Stratum 3
$n_1 = 20$	$n_2 = 8$	$n_3 = 12$
$N_1 = 155$	$N_2 = 62$	$N_3 = 93$
$W_1 = 0.5$	$W_2 = 0.2$	$W_3 = 0.3$
$f_1 = 0.129$	$f_2 = 0.129$	$f_3 = 0.129$
$\bar{y}_1 = 33.900$	$\bar{y}_2 = 25.125$	$\bar{y}_3 = 19.000$
$\hat{p}_1 = 0.8$	$\hat{p}_2 = 0.25$	$\hat{p}_3 = 0.50$
$s_1^2 = 35.358$	$s_2^2 = 232.411$	$s_3^2 = 87.636$

Example

- We want to estimate μ , the average number of hours of certification training in the year prior to the last certification.
- Estimate

$$\begin{aligned}\bar{y} &= \sum_h W_h \bar{y}_h \\ &= 0.5 \times 33.900 + 0.2 \times 25.125 + 0.3 \times 19.000 \\ &= 27.675 \text{ hours}\end{aligned}$$

Example

- Estimated variance

$$\begin{aligned}\widehat{\text{Var}}(\bar{y}) &= \sum_h W_h^2 \left(\frac{1 - f_h}{n_h} \right) s_h^2 \\ &= 0.5^2 \left(\frac{1 - 0.129}{20} \right) 35.358 \\ &\quad + 0.2^2 \left(\frac{1 - 0.129}{8} \right) 232.411 \\ &\quad + 0.3^2 \left(\frac{1 - 0.129}{12} \right) 87.636 \\ &= 1.97\end{aligned}$$

Example

- Estimated standard error $\sqrt{1.97} = 1.40$
- 95% CI using df formula

$$27.675 \pm t_{0.975, 21.1} 1.4034 = (24.757, 30.593)$$

- SAS uses $df = n - H = 37$

$$27.675 \pm t_{0.975, 37} 1.4034 = (24.831, 30.519)$$

Example in SAS

```
data all;
  input stratum id hours pass;
  cards;
1 1 35 1
1 2 28 1
1 3 26 1
.
.
.
.
.
.
3 11 34 0
3 12 24 1
;

data total;
  input stratum _TOTAL_;
  cards;
1 155
2 62
3 93
;
```

Example in SAS

```
proc surveymeans data=all total=total;  
  var hours;  
  strata stratum;
```

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Observations	40

Statistics

Variable	N	Mean	Std Error	95% CL for Mean	
			of Mean	Lower	Upper
hours	40	27.675000	1.403396	24.8314503	30.5185497

BIOS 662 Fall 2018

Survey Sampling, Part III

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

<https://sakai.unc.edu/portal>

Outline

- One-stage cluster sampling
- Systematic sampling
- Multi-stage cluster sampling
- Comments on cluster sampling
- Sampling overview

Cluster Sampling

- Partition the population into exhaustive and mutually exclusive *primary units* or *clusters*
- Each primary unit is composed of *secondary units*
- Select a sample of primary units using some sampling design (e.g., SRS)
- Record the y values of *every* secondary unit within the selected primary units

Cluster Sampling

- This seems similar to stratification, with cluster = strata
- However, these are different designs
- In stratification, we sample some units from every stratum
- In cluster sampling, we sample *all* secondary units from some clusters
- This is sometimes called *one-stage cluster sampling* or *single-stage cluster sampling*

Cluster Sampling Examples

- In a household survey for a small city, a probability sample of blocks is selected. Each block in this case represents a cluster of households. All households within selected blocks are surveyed.
- In a survey of first graders in the schools of a state, a probability sample of schools is selected. All first graders in a school would represent a cluster in this design.
- In a national sample of inpatient hospital visits for individuals with multiple sclerosis during some calendar year, a probability sample of hospitals is chosen. Each hospital in this instance represents a cluster of visits by patients with multiple sclerosis during that year.

Notation and Estimands

- N is the number of primary units in the population
- n is the number of primary units in the sample
- M_i is the number of secondary units in primary unit i
- The total number of secondary units in the population is

$$M = \sum_{i=1}^N M_i$$

Notation and Estimands

- y_{ij} is the value of the variable of interest for secondary unit j of primary unit i
- $y_i = \sum_j y_{ij}$
- Population total: $\tau = \sum_{i=1}^N y_i = \sum_i \sum_j y_{ij}$
- Population mean per primary unit: $\mu_p = \frac{\tau}{N}$
- Population mean per secondary unit: $\mu = \frac{\tau}{M}$
- Let $Z_i = 1$ if primary unit i is selected, 0 otherwise

Estimators

- Assume SRS of primary units/clusters, also known as *simple cluster sampling*
- An unbiased estimator of τ is:

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^N y_i Z_i = N\bar{y}$$

where $\bar{y} = \sum_i y_i Z_i / n$ is the sample mean of the primary unit totals

Estimators

- The variance of $\hat{\tau}$ is

$$\text{Var}(\hat{\tau}) = N(N - n) \frac{\sigma_u^2}{n}$$

where σ_u^2 is the finite population variance of the primary unit totals

$$\sigma_u^2 = \frac{1}{N - 1} \sum_{i=1}^N (y_i - \mu_p)^2$$

Estimators

- An unbiased estimator of the variance of $\hat{\tau}$ is

$$\widehat{\text{Var}}(\hat{\tau}) = N(N - n) \frac{s_u^2}{n}$$

where s_u^2 is the sample variance of the primary unit totals

$$s_u^2 = \frac{1}{n - 1} \sum_{i=1}^N (y_i - \bar{y})^2 Z_i$$

Estimators

- These results follow directly from the SRS derivations, thinking of the clusters as units in a population of size N with variables y_1, \dots, y_N
- An unbiased estimator of μ_p is given by $\bar{y} = \hat{\tau}/N$
- An unbiased estimator of μ is given by $\hat{\mu} = \hat{\tau}/M$
- Variances and unbiased estimators of variances follow accordingly

Cluster Sampling Principle

- Within-cluster variance does not affect variance of estimators
- Rather, it is only between-cluster variance that has an effect
- Thus, to minimize variance, clusters should be chosen to be as similar to one another as possible
- The ideal primary unit should be “representative”, that is, contain the full diversity of the population
(Thompson, page 118)
- This often runs counter to the practicalities of cluster sampling, e.g., where clusters are composed of geographically adjacent units

Systematic Sampling

- *Systematic Sampling*: A method of probability sampling in which elements on an ordered list are chosen by selecting elements a fixed distance apart on the list:
 1. Number units in sampling frame sequentially from 1 to N
 2. Choose a sampling interval k . If a sample of size about n is desired, k is usually the ratio, N/n , rounded to the nearest integer.
 3. Choose a random number between 1 and k . This is called the *random start* and will be denoted by g .
 4. Elements selected in the sample are those numbered g and every k^{th} element for the remainder of the list; that is, $g, g + k, g + 2k$, etc.

Systematic Sampling

- Systematic sampling can be viewed as a special form of cluster sampling.
- Specifically, the population can be viewed as consisting of k clusters each of which is a possible systematic sample which can be chosen. By choosing a random start and applying a fixed interval in selecting the sample, we are effectively randomly choosing one of the k possible clusters.
- Thus we can obtain an unbiased estimator of the population total or mean. However, because this sample contains just one cluster, it is not possible to obtain unbiased estimators of the variances.

Multi-Stage Cluster Sampling

- *Multi-Stage Cluster Sampling:* A method of probability sampling in which the sample of elements is chosen in two or more stages. Second stage sampling units are chosen from the sampling units selected in the first stage. Third stage units are chosen from second stage sampling units; and so forth.

- Example: Household sample of the non-institutionalized population in Virginia

Primary Sampling Units: Minor Civil Divisions

Secondary Sampling Units: Small Groups of Blocks

Tertiary Sampling Units: Households

Multi-Stage Cluster Sampling

- Example: National Sample of Hospital Discharges
 - PSU: Small Groups of Counties
 - SSU: Hospitals
 - TSU: Patient Medical Records
- Example (Tate and Hudgens, *AJE*, 2007): Estimating the number of individuals at high risk for HIV in Osh, Kyrgyzstan
 - PSU: Public venues within the city where risky sexual and drug-use behaviors occur
 - SSU: Individuals socializing at these venues

Two-Stage Cluster Sampling

- We consider a two-stage design with SRS at each stage
- First stage: SRS of n primary units selected
- Second stage: SRS of m_i secondary units selected from the i^{th} selected primary unit, for $i = 1, \dots, n$
- $\mu_i = y_i/M_i$ is the mean per secondary unit in the i^{th} primary unit
- Z_i is as before
- $Z_{ij} = 1$ if the j^{th} secondary unit of the i^{th} primary unit is in the sample, 0 otherwise

Two-Stage Cluster Sampling

- If the i^{th} primary unit is selected, an estimator of the total y -value for that unit (that is, y_i) is

$$\hat{y}_i = \frac{M_i}{m_i} \sum_{j=1}^{M_i} y_{ij} Z_{ij} = M_i \bar{y}_i$$

where

$$\bar{y}_i = \frac{1}{m_i} \sum_{j=1}^{M_i} y_{ij} Z_{ij}$$

- Because SRS is used at the second stage, this estimator is conditionally unbiased

$$E(\hat{y}_i | Z_i = 1) = y_i$$

Two-Stage Cluster Sampling

- An unbiased estimator of the population total is given by

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^N \hat{y}_i Z_i$$

- To prove this, we use the fact that

$$E(\hat{\tau}) = E\left(E(\hat{\tau}|Z_1, \dots, Z_n)\right)$$

Two-Stage Cluster Sampling

- First evaluate the inner expectation

$$\begin{aligned} E(\hat{\tau}|Z_1, \dots, Z_n) &= E\left(\frac{N}{n} \sum_{i=1}^N \hat{y}_i Z_i \middle| Z_1, \dots, Z_N\right) \\ &= \frac{N}{n} \sum_{i=1}^N E(\hat{y}_i | Z_i = 1) Z_i \\ &= \frac{N}{n} \sum_{i=1}^N y_i Z_i \end{aligned}$$

- Then evaluate the outer expectation

$$E(\hat{\tau}) = E\left(\frac{N}{n} \sum_{i=1}^N y_i Z_i\right) = \frac{N}{n} \sum_{i=1}^N y_i E(Z_i) = \sum_{i=1}^N y_i = \tau$$

Two-Stage Cluster Sampling

- The variance of $\hat{\tau}$ is

$$\text{Var}(\hat{\tau}) = N(N - n) \frac{\sigma_u^2}{n} + \frac{N}{n} \sum_{i=1}^N M_i(M_i - m_i) \frac{\sigma_i^2}{m_i}$$

where

$$\sigma_u^2 = \frac{1}{N - 1} \sum_{i=1}^N (y_i - \mu_p)^2$$

(as before) and

$$\sigma_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \mu_i)^2$$

for $i = 1, \dots, N$

Two-Stage Cluster Sampling

- Note that the first term in $\text{Var}(\hat{\tau})$ is the variance that would be obtained if every secondary unit in a selected primary unit is observed (that is, $m_i = M_i$ for all i). So the second term can be viewed as a penalty for having to estimate y_i .
- Similarly, note that the second term equals the $\text{Var}(\hat{\tau})$ when $n = N$, that is, every primary unit is selected. In this case, we recover the variance from stratified sampling. So the first term can be viewed as a penalty for using cluster sampling instead of stratified sampling.

Two-Stage Cluster Sampling

- To derive $\text{Var}(\hat{\tau})$, we will use the fact

$$\text{Var}(\hat{\tau}) = \text{Var}\left(E(\hat{\tau}|Z_1, \dots, Z_N)\right) + E\left(\text{Var}(\hat{\tau}|Z_1, \dots, Z_N)\right)$$

- For the first term, we have

$$\text{Var}\left(E(\hat{\tau}|Z_1, \dots, Z_N)\right) = \text{Var}\left(\frac{N}{n} \sum_{i=1}^N y_i Z_i\right) = N(N-n) \frac{\sigma_u^2}{n}$$

where the second equality follows from results for SRS

- To evaluate the second term, first note that

$$\text{Var}(\hat{y}_i|Z_i = 1) = M_i(M_i - m_i) \frac{\sigma_i^2}{m_i}$$

Two-Stage Cluster Sampling

- Therefore

$$\begin{aligned}\text{Var}(\hat{\tau}|Z_1, \dots, Z_n) &= \text{Var}\left(\frac{N}{n} \sum_{i=1}^N \hat{y}_i Z_i \middle| Z_1, \dots, Z_N\right) \\ &= \left(\frac{N}{n}\right)^2 \sum_{i=1}^N \text{Var}(\hat{y}_i | Z_i = 1) Z_i \\ &= \left(\frac{N}{n}\right)^2 \sum_{i=1}^N M_i(M_i - m_i) \frac{\sigma_i^2}{m_i} Z_i\end{aligned}$$

- Thus

$$E(\text{Var}(\hat{\tau}|Z_1, \dots, Z_n)) = \frac{N}{n} \sum_{i=1}^N M_i(M_i - m_i) \frac{\sigma_i^2}{m_i}$$

Two-Stage Cluster Sampling

- An unbiased estimator of the variance of $\hat{\tau}$ is

$$\widehat{\text{Var}}(\hat{\tau}) = N(N - n) \frac{s_u^2}{n} + \frac{N}{n} \sum_{i=1}^N M_i(M_i - m_i) \frac{s_i^2}{m_i} Z_i$$

where

$$s_u^2 = \frac{1}{n-1} \sum_{i=1}^N (\hat{y}_i - \hat{\mu}_p)^2 Z_i$$

and

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_i)^2 Z_{ij}$$

for $i = 1, \dots, N$

- The proof is left as an exercise

Two-Stage Cluster Sampling

- Estimators for population means follow immediately:

$$\hat{\mu}_p = \hat{\tau}/N \text{ is unbiased for } \mu_p$$

$$\hat{\mu} = \hat{\tau}/M \text{ is unbiased for } \mu$$

- Variance expressions follow from $\text{Var}(\hat{\tau})$ divided by the appropriate constant

Two-Stage Cluster Sampling: Example

- SRS of $n = 3$ primary units selected from a population of $N = 100$ primary units
- For each of the selected primary units, SRS of $m_i = 2$ secondary units selected
- Sizes of the three selected primary units: 24, 20, 15
- Y -values for the first selected primary unit: 8, 12
- Y -values for the second selected primary unit: 0, 0
- Y -values for the third selected primary unit: 1, 3

Two-Stage Cluster Sampling: Example

- Estimate of the population total:

$$\hat{\tau} = \frac{100}{3} \left(24 \cdot \frac{8+12}{2} + 20 \cdot \frac{0+0}{2} + 15 \cdot \frac{1+3}{2} \right) = 9,000$$

- Estimate of the mean per primary unit:

$$\hat{\mu}_p = \frac{\hat{\tau}}{N} = 90$$

- Sample variance across primary unit totals:

$$s_u^2 = \frac{1}{3-1} \left((240 - 90)^2 + (0 - 90)^2 + (30 - 90)^2 \right) = 17,100$$

Two-Stage Cluster Sampling: Example

- After computing the sample variances within the selected primary units, we have

$$\begin{aligned}\widehat{\text{Var}}(\hat{\tau}) &= 100(100 - 3) \frac{17100}{3} \\ &\quad + \frac{100}{3} \left(24(24 - 2) \frac{8}{2} + 20(20 - 2) \frac{0}{2} + 15(15 - 2) \frac{2}{2} \right) \\ &= 55,366,900\end{aligned}$$

Comments on Cluster Sampling

- Simple cluster sampling is epsem
- One-stage cluster sampling generally yields estimates with relatively larger variances (i.e., lower precision) than samples of the same size which are chosen by (individual) element (i.e., non-cluster) sampling. The amount of the increase in variance is directly related to the average sample cluster size.

Comments on Cluster Sampling

- Because units of clusters are often close in geographic proximity, the average cost per sample element can be reduced substantially over individual element sampling if cluster sampling is used
- The size of the cost reduction is directly related to the average size of the clusters that are used
- Elements in a cluster are usually similar (i.e., clusters are internally homogeneous), so the amount of information gathered by the survey may not be increased substantially as additional units are surveyed within a cluster
- So sample cluster sizes should not be too large

Comments on Cluster Sampling

- As a general rule, the number of clusters in the population should be large which means that the average size of clusters should be kept as small as possible.
- The survey statistician frequently has some choice in the size of clusters that are used in a survey
- In making this choice, the cost advantages of large (sample) clusters must be properly weighed against the statistical advantages of smaller (sample) clusters

Comments on Cluster Sampling

- Cluster sampling eliminates the need for a sampling frame consisting of a list of all elements in the population
- Because clusters are the units being sampled, a listing of all clusters in the population constitutes an appropriate frame
- Through multi-stage cluster sampling, most of the cost savings can be retained while gaining back some of the statistical losses (i.e., larger variances) of one-stage cluster sampling

Summary

- Identified several basic sampling designs (on next slide)
- Derived properties (expectations, variances, ...) of various estimators
- Illustrated with real data sets
- 664 [164] SAMPLE SURVEY METHODOLOGY (STAT 358) (3). Prerequisite, BIOS 550 or equivalent or permission of the instructor. Fundamental principles and methods of sampling populations, with primary attention given to simple random sampling, stratified sampling, and cluster sampling. Also, the calculation of sample weights, dealing with sources of nonsampling error, and analysis of data from complex sample designs are covered. Practical experience in sampling is provided by student participation in the design, execution, and analysis of a sampling project. Spring.

Summary

- SRS
- Stratified
 - *Proportionate* – default; always better than SRS
 - *Optimal, disproportional, balanced*
- Cluster
 - *One stage* – SRS of clusters; sample all within cluster
 - *Systematic (list)*
 - *Multi-stage*, for example, blocks then dwellings