# 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.

  $\rightarrow$ 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.

  $\rightarrow$ 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.

$\rightarrow$ 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)

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)`

```
> 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
design1
design2
power
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
effect size

signal intensity
(Lee et al. 2007)
1.0
0.5
0.0
−1.0
signal intensity
(this study)
0.5
0.0
−0.5
−1.0
R=1.05
A=−1.05
RVS167
SAC7
RVS167
SAC7
1250000
1252000
chr4 (bp)
signal intensity
(Lee et al. 2007)
signal intensity
(this study)
0.2
−0.2
−0.6
R=1.13
A=−1.02
RPS17B
ADA2
ECM11
RPS17B
ADA2
UTP6
1354000
1356000
chr4 (bp)

# 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:

$$x_1, x_2, \ldots, x_n$$

- Mean:

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

- 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

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

then

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

- If

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

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: $x_1, x_2, \ldots, x_n$
- Order data from smallest to largest

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

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

$$x_{(1)} = \min\{x_1, x_2, \ldots, x_n\}$$
$$x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$$

- $(1), (2), \ldots, (n)$ 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

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

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: $P$th *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:

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

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

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

- Let $y_i = \log(x_i)$ for $i = 1, 2, \ldots, n$. 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 $\sigma^2$, 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

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

- Note

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

is called the *interquartile range*

# Comments

- $\hat{\xi}_{0.25}$ is the lower quartile
- $\hat{\xi}_{0.75}$ is the upper quartile
- Epidemiologists often use quantiles to categorize a continuous variable
- Splitting a variable at $\hat{\xi}_{0.25}, \hat{\xi}_{0.5}, \hat{\xi}_{0.75}$ 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

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

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

  $\bar{x} = 2.2$

  mode $= 2$

  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 | Generation 2nd |
|---|---|---|---|
| $\leq 104$ | 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 | Generation 2nd |
|---|---|---|---|
| $\leq 104$ | 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 | Generation 2nd |
|---|---|---|---|
| $\leq 104$ | 0.098 | 0.015 | 0.015 |
| $\leq 114$ | 0.220 | 0.103 | 0.100 |
| $\leq 124$ | 0.371 | 0.289 | 0.285 |
| $\leq 134$ | 0.533 | 0.414 | 0.507 |
| $\leq 144$ | 0.668 | 0.570 | 0.729 |
| $\leq 154$ | 0.785 | 0.715 | 0.858 |
| $\leq 164$ | 0.859 | 0.802 | 0.928 |
| $< \infty$ | 1.000 | 1.000 | 1.000 |
| Total | 2232 | 263 | 1561 |

# ECD – Example

Native
First
Second
ECD
BP
0.0
0.2
0.4
0.6
0.8
1.0
110
120
130
140
150
160

# 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 | 0012344799033344566778
  30 | 000011112233344566677788800112222344555678899 99
  32 | 00000122333445666666999011112334444445555555555566666777789
  34 | 0000000111222223333344455666667888888999000000001111222233344445555+5
  36 | 000000001111222333333444455555556666677778889999900000002233344445556+2
  38 | 00000111223333334555555677999000122333344455678888999 9
  40 | 0011133446788888999990034566789 99
  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 | 0012344799033344566778
  30 | 000011112233344566677788800112222344555678899 99
  32 | 00000122333445666666999011112334444445555555555566666777789
  34 | 0000000111222223333344455666667888888999000000001111222233344445555566666677777788889
  36 | 000000001111222333333444455555555666677778889999900000002233344445556666667777789 99
  38 | 00000111223333334555555677999000122333344455678888999 9
  40 | 0011133446788888999990034566789 99
  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 | 000011112233344566677788
  31 | 0011222234455567889999
  32 | 0000012233344566666999
  33 | 01111233444445555555555566666777789
  34 | 00000001112222233333444556666678888889999
  35 | 000000001111222233344445555666666777777788889
  36 | 0000000011112223333334444555555566667777888999999
  37 | 00000002233344445556666667777789999
  38 | 000001112233333345555555677999
  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 x IQR; smallest observation $\geq \hat{\zeta}_{0.25} -$ 1.5 x 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;
```

```
                      Histogram                        #  Boxplot
    4.7+*                                              1     0
       .**                                             3     |
       .*****                                          9     |
       .****************                              32     |
       .****************************                  55     |
       .****************************************      82  +-----+
       .*******************************************   85  *--+--*
    3.3+******************************                59  +-----+
       .************************                      47     |
       .***********                                   22     |
       .*****                                          9     |
       .****                                           7     0
       .***                                            5     0
       .*                                              1     0
    1.9+*                                              1     0
        ----+----+----+----+----+----+----+----+---
        * may represent up to 2 counts
```

# 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

$$x_i = ax_{i-1} (\text{mod } m),\ i = 1, 2, 3, \ldots$$

  - 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 $\{1, 2, 3, \ldots, (m-1)\}$
  - 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 $U(\texttt{min},\texttt{max})$
  - 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 $x$: $\quad x = -\frac{1}{\lambda}\ln(1-u)$.

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

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 $p^{\text{th}}$ quantile of $X$, say $\zeta_p$, 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

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

- $E(g(X)) = \sum_x g(x) p_X(x)$ (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:

  $\mu$ population mean; $\sigma^2$ 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 $\mu$ and a population variance $\sigma^2$, the sampling distribution of the mean ($\bar{Y}$) has mean $\mu$ and variance $\sigma^2/n$
- Definition 4.16. The standard deviation of the sampling distribution is called the *standard error*
- For example, the standard error of $\bar{Y}$ is $\sigma/\sqrt{n}$

# 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:

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

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

# Normal Distribution

# Standard Normal Distribution

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

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

- 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 $\mu$ 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

$$E(Y) = aE(X) + b$$
$$\mathrm{Var}(Y) = a^2\,\mathrm{Var}(X)$$

- 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

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

$$= \Phi(\tfrac{b-\mu}{\sigma}) - \Phi(\tfrac{a-\mu}{\sigma})$$

# 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 $\mu = 16$ mmHg
  and variance $\sigma^2 = 9$ mmHg$^2$
- If pressure greater than 20 mmHg is considered abnormal, what proportion of the population is abnormal?

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

## Example (continued)

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

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

# 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 $\Pr[\mu_X - \sigma_X \le X \le \mu_X + \sigma_X] \approx 0.54$

## Some Approximations

- The following holds for any random variable $Y$ with mean $\mu$ and variance $\sigma^2$

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

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

- For example, if $K = 2$,

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

  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 $Y_1, Y_2, \ldots, Y_n$ 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 $n \to \infty$.

# Central Limit Theorem (CLT)

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

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

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

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

# Confidence Interval for $\mu$

- Suppose $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$

- 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 $\mu$

# Confidence Interval for $\mu$

- 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 $\mu$

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

# Confidence Interval for $\mu$

- $100(1-\alpha)\%$ CI for $\mu$

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

or

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

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

| $100(1-\alpha)\%$ | $\alpha$ | $z_{1-\alpha/2}$ |
|---|---|---|
| 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 $\mu$
- If we draw 100 different random samples, on average $100(1 - \alpha)\%$ of them will contain $\mu$

- To decrease the width of CI:
  - increase $\alpha$, 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 $\sigma^2$ 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 $\mu$ and variance $\sigma^2$, 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 $\chi^2_{n-1}$

# $\chi^2$ Distribution

# $t$ Distribution

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

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

  will follow the $t$-distribution with $\nu$ degrees of freedom.

- We know

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

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

# $t$ Distribution

# CI for $\mu$ when $\sigma^2$ 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 $\mu$ 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 $\mu$:

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

- 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 $\sigma$

# 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 $Y_1, \ldots, Y_n$ is a random sample from a distribution with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$ for $i = 1, \ldots, n$, 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 $\mu$ requires knowledge of $\sigma^2$
- To use the CLT when $\sigma^2$ 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

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

# 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 $n \to \infty$

# Large Sample CI for $\mu$

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

| Normal | $\sigma^2$ known | $n$ large | Confidence Interval |
|---|---|---|---|
| $\surd$ | $\surd$ | – | $\bar{Y} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$ |
| – | $\surd$ | $\surd$ | $\bar{Y} \pm z_{1-\alpha/2}(\sigma/\sqrt{n})$ |
| $\surd$ | – | – | $\bar{Y} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n})$ |
| – | – | $\surd$ | $\bar{Y} \pm z_{1-\alpha/2}(s/\sqrt{n})$ |
| – | – | – | 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 $\mu$ is

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

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 $\{x_1, \ldots, x_n\}$; call this $\boldsymbol{x}^*(1)$
2. Compute $Z^*(1)$ as described on the next slide
3. Do steps 1 and 2 a total of $B$ times, to obtain $Z^*(1), Z^*(2), \ldots, Z^*(B)$
4. Let $\hat{t}_{(\alpha/2)}$ be the $\alpha/2$ sample quantile of $\{Z^*(1), \ldots, Z^*(B)\}$; similarly for $\hat{t}_{(1-\alpha/2)}$

- The order of steps 2 and 3 can be interchanged, first obtaining $B$ bootstrap samples $\boldsymbol{x}^*(1), \ldots, \boldsymbol{x}^*(B)$ and then computing $Z^*(b)$ for each bootstrap sample

# Bootstrap t-intervals

- Step 2. For each bootstrap sample compute

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

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

$$\widehat{se}^*(b) = \sqrt{\mathrm{Var}\{\boldsymbol{x}^*(b)\}/n}$$

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

# 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 $\mu$
- 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

| $n$ | Population distribution | $t$ | Bootstrap $t$ | "Simple" bootstrap |
|---|---|---|---|---|
| 20 | N(1,1) | 0.949 | 0.946 | 0.924 |
| | Chi-squared (1 df) | 0.891 | 0.936 | 0.876 |
| | Chi-squared (1 df)* | 0.905 | 0.942 | 0.875 |
| | Exponential(1) | 0.922 | 0.945 | 0.902 |
| 25 | N(1,1) | 0.949 | 0.947 | 0.931 |
| | Chi-squared (1 df) | 0.922 | 0.944 | 0.889 |
| | Exponential(1) | 0.902 | 0.939 | 0.889 |

* 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 $X_1, \ldots, X_n$ 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

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

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

# Binomial Random Variable

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

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

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

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

# 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 |
| $\vdots$ | |

- 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 Distribution Free | | -------Order Statistics------- 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 $n \to \infty$

# Large Sample CI for Any Quantile

- In general,

$$\Pr[\zeta_p < X_{(r)}] = \sum_{i=0}^{r-1} \binom{n}{i} F(\zeta_p)^i \{1 - F(\zeta_p)\}^{n-i}$$

$$= \sum_{i=0}^{r-1} \binom{n}{i} p^i q^{n-i}$$

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

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

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

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

- 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 $\zeta_p$ is given by

$$(X_{(\lfloor r \rfloor)}, X_{(\lceil s \rceil)})$$

- Note: $n$ large enough ensures $\lfloor r \rfloor, \lceil s \rceil \in \{1, \ldots, n\}$

# 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 $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean $\mu$ and variance $\sigma^2$
- Recall (result 4.4 on p. 95 of the text)

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

- Therefore

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

- Implying

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

# CI for Variance

- Because the $\chi^2$ distribution is not symmetric, we need to look up both $\chi^2_{\alpha/2,n-1}$ and $\chi^2_{1-\alpha/2,n-1}$
- 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 $\sigma^2$:

$$(22(3701.36)/36.78, 22(3701.36)/10.98)$$
$$= (2213.973, 7416.203)$$

- 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

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

where $g_2 = a_4 - 3$ and $a_4$ is an estimate of $\alpha_4$ (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 $\theta_0$ 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 $\theta$, 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 $\alpha$ such that:

  If the computed probability is less than or equal to $\alpha$, we reject $H_0$
- $\alpha$ 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 $\alpha$ and under $F_0$ determine a *critical region* or *rejection region* $C_\alpha$ such that

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

- Values at the boundaries of $C_\alpha$ 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 $\alpha$ and establish $C_\alpha$
5. Carry out study and collect data
6. Compute statistic from data
7. If statistic is in $C_\alpha$, 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 $\theta$ denote the mean BP change after 3 weeks
- Hypotheses

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

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

# Example cont.

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

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

  for some constant $c$

- In particular, we want $c$ such that

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

# 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

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

that is, if

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

- 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            76.4444444

           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 $H_0$ | $\surd$ | Type II |
| | Reject $H_0$ | Type I | $\surd$ |

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

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

- 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 $\alpha$ 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, $Y_1, \ldots, Y_n$, 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 Mean | Mean | Upper CL 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)

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

and Slutsky's theorem

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

- Approximation improves as $n \to \infty$
- 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

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

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

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

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

- 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 $\mu$

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

- Can show: Reject $H_0$ if and only if the CI excludes $\mu_0$

# Testing/Estimation: Large Sample

# Testing/Estimation: Large Sample

- Theorem: Reject $H_0$ if and only if CI excludes $\mu_0$
- Sketch of proof: Suppose CI excludes $\mu_0$, 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 $Y_1, \ldots, Y_n$ are iid continuous from $F$ with median $\zeta_{0.5}$
- Hypotheses

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

- 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 $H_0$

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

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

$$= \frac{1}{2^n} \sum_{i=0}^{r} \binom{n}{i}$$

# Sign Test

- Critical region

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

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$(n = 10, \pi = 0.5)$

| r | Cumulative 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            76.4444444
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 $n \to \infty$

# 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

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

$$= \Pr\left[Z \leq \frac{r - (n-1)/2}{\sqrt{n/4}}\right]$$

# 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 $Y_1, Y_2, \ldots, Y_n$ 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 $\left|Y_i'\right|$ 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 | $Y_i$ diff | $\lvert Y_i \rvert$ absol | rank | sign*rank |
|---|---|---|---|---|---|---|---|
| 1. | calcium | 107 | 100 | −7 | 7 | 6 | −6 |
| 2. | calcium | 110 | 114 | 4 | 4 | 4 | 4 |
| 3. | calcium | 123 | 105 | −18 | 18 | 10 | −10 |
| 4. | calcium | 129 | 112 | −17 | 17 | 9 | −9 |
| 5. | calcium | 112 | 115 | 3 | 3 | 3 | 3 |
| 6. | calcium | 111 | 116 | 5 | 5 | 5 | 5 |
| 7. | calcium | 107 | 106 | −1 | 1 | 1 | −1 |
| 8. | calcium | 112 | 102 | −10 | 10 | 7 | −7 |
| 9. | calcium | 136 | 125 | −11 | 11 | 8 | −8 |
| 10. | calcium | 102 | 104 | 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 | | | | |
|---|---|---|---|---|
| 1 | 2 | 3 | 4 | $S^+$ |
| − | − | − | − | 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

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

- 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: $\boldsymbol{Y} = \{23, 25, 45, 13, 23, 46\}$

  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

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

  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

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

# 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

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

- 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

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

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

# 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 $\mu_1$ and $\mu_2$ and variances $\sigma_1^2$ and $\sigma_2^2$, 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 $\mu_1$ and $\mu_2$ 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.

- $\bar{y}_1 = 7.08, \quad s_1 = 0.899$
- $\bar{y}_2 = 6.26, \quad s_2 = 0.961$
- 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

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

- We know that (assuming normality)

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

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

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

# F Distribution

# Homogeneity of Variance

- Let

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

- 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

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

- Thus

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

# 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 $H_0$, $Z \dot{\sim} N(0, 1)$
- Approximation gets better as $n_1, n_2 \to \infty$
- 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

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

- Note 5.2 of text:

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

# Welch-Satterthwaite Approximation

- Welch (Biometrika 1938), SAS and R

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

- Use $\lfloor \mathrm{df} \rfloor$ if using tables

# Welch-Satterthwaite Approximation: Example

- Premature birth example

$$n_1 = n_2 = 15$$

$$s_1 = 0.8994,\ s_2 = 0.9605$$

$$\mathrm{df}_{\mathrm{text}} = 29.86$$

$$\mathrm{df}_{\mathrm{welch}} = 27.88$$

# 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 |
|---|---|---|---|---|
| $\surd$ | $\surd$ | | | $\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$ |
| | $\surd$ | | $\surd$ | |
| $\surd$ | | $\surd$ | | $\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ |
| | | | $\surd$ | $\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim N(0,1)$ |
| $\surd$ | | | | $\frac{(\bar{Y}_1-\bar{Y}_2)-\delta}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{df}$ |
| | | | | 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 $Y_{1j}, \ldots, Y_{n_j j}$ 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 $\Delta$ 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:

$$\text{control} \;\; (n_1 = 3) \;\; \text{or drug} \;\; (n_2 = 4)$$

- 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

$$\mathrm{Var}(W_1) = \frac{n_1 n_2 (N+1)}{12}$$

(cf. Lehmann 1998, Example 3, p 332)

# Large Sample Approximation

- If $n_1$ and $n_2$ are large

$$Z = \frac{W_1 - E(W_1)}{\sqrt{\mathrm{Var}(W_1)}}$$

will be approximately $N(0,1)$

- Approximation is good for $n_1, n_2 \geq 12$
- If there are ties

$$\mathrm{Var}(W_1) = \frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2}{12N(N-1)} \sum_{i=1}^{q} t_i(t_i - 1)(t_i + 1)$$

# 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:

$q = 7;\ \ t_1 = t_2 = 2;\ \ t_3 = t_4 = 3;\ \ t_5 = t_6 = t_7 = 2$

$$\sum_{i=1}^{q} t_i(t_i - 1)(t_i + 1) = 78$$

$$\mathrm{Var}(W_1) = 581.25 - \frac{78(15)^2}{12(30)(29)} = 579.57$$

# 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

                   Sum of     Expected      Std Dev         Mean
trt       N        Scores     Under H0     Under H0        Score
-----------------------------------------------------------------
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

$$\Pr\left[|W_1 - E(W_1)| \geq |w_1 - E(W_1)|\right]$$

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 $\boldsymbol{Y}_1 = (65, 70, 73)$ and $\boldsymbol{Y}_2 = (70, 89)$
- There are $\binom{5}{2} = 10$ possible rankings for group 1

| Ranks | $W_1$ | $\lvert W_1 - E(W_1)\rvert$ | Ranks | $W_1$ | $\lvert W_1 - E(W_1)\rvert$ |
|---|---|---|---|---|---|
| 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

                      Sum of      Expected       Std Dev          Mean
group       N         Scores      Under H0      Under H0         Score
-----------------------------------------------------------------------
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

$$(Y_{1i}, Y_{2j});\ i = 1, 2, \ldots, n_1;\ j = 1, 2, \ldots, n_2$$

- 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 $\Delta$.
- 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 $\Delta$
- Estimate $\Delta$ 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 $Y_{1i} > Y_{2j} + \hat{\Delta}$ half of the time
- Thus

$$\hat{\Delta} = \text{median}\{Y_{1i} - Y_{2j} : i = 1, \ldots, n_1; j = 1, \ldots, n_2\}$$

- CI for $\Delta$?

# 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 $\alpha$ 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 $\theta$
- 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 $\theta$ for which we would not reject $H_0$ at the $\alpha$ level of significance

- For example, consider the Wilcoxon Rank Sum test:

  For a given value of $W_1$, find all values of $\Delta$ 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 | Group2 | $\bar{Y}_1 - \bar{Y}_2$ | 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: $\boldsymbol{Y}_1 = (1, 2, 4)$, $\boldsymbol{Y}_2 = (3, 5)$

# Kolmogorov-Smirnov Test: Example

# Kolmogorov-Smirnov Test: Example cont.

- Observe $d = 2/3$
- There are 10 possible group assignments

| $\boldsymbol{Y}_1$ | $D$ | $\boldsymbol{Y}_1$ | $D$ |
|---|---|---|---|
| 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

                      EDF at    Deviation from Mean
group       N       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

- $X_1, \ldots, X_n \sim \text{Bernoulli}(\pi)$
- $Y = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, \pi)$
- Four key conditions

  1. Binary response (0/1)
  2. Observed a known number of times $n$
  3. Success probability $(\pi)$ 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 |
| $\vdots$ | $\vdots$ | |
| 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 $n \to \infty$
- 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 $\pi_0$ such that $H_0 : \pi = \pi_0$ would not be rejected
- To get an exact $100(1-\alpha)\%$ CI for $\pi$, solve these equations for $\pi_L$ and $\pi_U$:

$$\Pr[Y \geq y | \pi = \pi_L] = \sum_{k=y}^{n} \binom{n}{k} \pi_L^k (1-\pi_L)^{n-k} = \alpha/2$$

$$\Pr[Y \leq y | \pi = \pi_U] = \sum_{k=0}^{y} \binom{n}{k} \pi_U^k (1-\pi_U)^{n-k} = \alpha/2$$

- 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 $X_1, X_2, \ldots, X_n$ 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 $\pi$ 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 $\pi$ 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
  - $\chi^2$ test

# Comparing Two Proportions

- Put the data in a $2 \times 2$ 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

$$\Pr[n_{11} = k | m_1, n_1, n_2] = \frac{\binom{n_1}{k}\binom{n_2}{m_1-k}}{\binom{N}{m_1}}$$

$$= \frac{n_1!\, n_2!\, m_1!\, m_2!}{N!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}$$

- 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 $2 \times 2$ 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 $i^{\text{th}}$ sample; $i = 1, 2$
- Estimator of $\pi_i$ 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 $\pi_i$ 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 $\pi_i$ 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 $\pi_1$ denote the probability of OC use given a case (MI) and $\pi_2$ 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

$$p_1 = 29/234 = 0.124; \;\; p_2 = 135/1742 = 0.078$$

- 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: $\chi^2$ Test

- Alternative test of $H_0 : \pi_1 = \pi_2$ is the $\chi^2$ test
- Recall $2 \times 2$ 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

$$X^2 = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_1 n_2 m_1 m_2} \sim \chi^2_1$$

- Critical region for $H_A : \pi_1 \neq \pi_2$

$$C_\alpha = \{X^2 : X^2 \geq \chi^2_{1,1-\alpha}\}$$

# Comparing Two Proportions: $\chi^2$ 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: $\chi^2$ 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

$$X^2 = \frac{1976(29 \times 1607 - 135 \times 205)^2}{234 \times 1742 \times 1812 \times 164} = 5.84$$

# $\chi^2$ 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
          ------------------------------------------------------
          Chi-Square                     1      5.8443    0.0156
```

# $\chi^2$ 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: $\chi^2$ Test

- Note: $\sqrt{5.84} = 2.42$ and $\sqrt{3.84} = 1.96$
- Intuition: If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$
- Indeed, for 2-sided tests, the $\chi^2$ 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 $\chi^2$ or $Z$ test, i.e.,

$$\begin{aligned} C_\alpha &= \{X^2 : X^2 > \chi^2_{1,1-\alpha}\} \\ \text{or} \quad C_\alpha &= \{z : |z| > z_{1-\alpha/2}\} \end{aligned}$$

- For large samples and $H_A : \pi_1 < \pi_2$ or $H_A : \pi_1 > \pi_2$, use $Z$ test, i.e.,

$$\begin{aligned} C_\alpha &= \{z : z < -z_{1-\alpha}\} \\ \text{or} \quad C_\alpha &= \{z : z > z_{1-\alpha}\} \end{aligned}$$

# 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 $2 \times 2$ 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

$$\begin{aligned} \pi_1 &= \Pr[\text{ disease} \mid \text{exposed }] \\ \text{and} \quad \pi_2 &= \Pr[\text{ disease} \mid \text{not exposed }] \end{aligned}$$

- Risk difference:

$$\text{RD} = \pi_1 - \pi_2$$

- Risk ratio (relative risk):

$$\text{RR} = \pi_1/\pi_2$$

- Odds ratio (cross product ratio):

$$\text{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

## Measures of Association: Estimands

- Independence or no association corresponds to

$$\mathrm{RR} = 1 \quad \text{and} \quad \mathrm{OR} = 1$$

- $\mathrm{OR}, \mathrm{RR} \in [0, \infty)$
- $\mathrm{RR} = 4$ implies an exposed person is 4 times as likely to have the disease as an unexposed person
- $\mathrm{OR} = 4$ implies the odds of disease in the exposed is 4 times that in the unexposed

# Measure of Association: Estimands

- Note

$$\text{OR/RR} = \left[\frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}\right] \Big/ \left[\frac{\pi_1}{\pi_2}\right] = \frac{1-\pi_2}{1-\pi_1}$$

- If disease rare,

$$1-\pi_1 \approx 1-\pi_2 \approx 1$$

- In this case, OR $\approx$ 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:

$$\widehat{\mathrm{RD}} = p_1 - p_2 = (n_{11}/n_1) - (n_{21}/n_2)$$

- Relative risk:

$$\widehat{\mathrm{RR}} = p_1/p_2 = (n_{11}/n_1)/(n_{21}/n_2)$$

- Odds ratio:

$$\widehat{\mathrm{OR}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

# Estimating RR in Case-Control Studies

- In case-control studies, $\widehat{\mathrm{RR}}$ 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 | |
|---|---|---|---|
| Exposed | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1\cdot}$ |
| Unexposed | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2\cdot}$ |
| | $\pi_{\cdot 1}$ | $\pi_{\cdot 2}$ | |

# 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 | $\frac{m_1\pi_{11}}{\pi_{\cdot 1}}$ | $\frac{m_2\pi_{12}}{\pi_{\cdot 2}}$ |
| Unexposed | $\frac{m_1\pi_{21}}{\pi_{\cdot 1}}$ | $\frac{m_2\pi_{22}}{\pi_{\cdot 2}}$ |
| | $m_1$ | $m_2$ |

# Estimating RR in Case-Control Studies

- Therefore

$$\widehat{\mathrm{RR}} \approx \frac{\left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}}\right) / \left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}} + \frac{m_2\pi_{12}}{\pi_{\cdot 2}}\right)}{\left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}}\right) / \left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{22}}{\pi_{\cdot 2}}\right)}$$

$$= \frac{\pi_{11} \times \left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{22}}{\pi_{\cdot 2}}\right)}{\pi_{21} \times \left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}} + \frac{m_2\pi_{12}}{\pi_{\cdot 2}}\right)}$$

$$= \frac{\frac{m_1\pi_{11}\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{11}\pi_{22}}{\pi_{\cdot 2}}}{\frac{m_1\pi_{11}\pi_{21}}{\pi_{\cdot 1}} + \frac{m_2\pi_{12}\pi_{21}}{\pi_{\cdot 2}}}$$

- This depends on the choice of $m_1$ and $m_2$; for instance, $\widehat{\mathrm{RR}} \to 1$ as $m_1 \to \infty$ for fixed $m_2$

# Estimating RR in Case-Control Studies

- On the other hand, we would expect

$$\widehat{\mathrm{OR}} \approx \frac{\left(\frac{m_1\pi_{11}}{\pi_{\cdot 1}}\right) / \left(\frac{m_2\pi_{12}}{\pi_{\cdot 2}}\right)}{\left(\frac{m_1\pi_{21}}{\pi_{\cdot 1}}\right) / \left(\frac{m_2\pi_{22}}{\pi_{\cdot 2}}\right)}$$

$$= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}$$

- For a rare disease, $\pi_{11}$ and $\pi_{21}$ 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 $\widehat{\mathrm{OR}} \approx \mathrm{RR}$ in this case.

# Estimating OR in Case-Control Studies

- Intuitively, why does $\widehat{\mathrm{OR}}$ estimate OR in a case-control study?

$$\mathrm{OR} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}$$

$$= \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\frac{\pi_{11}}{\pi_{11}+\pi_{21}} \Big/ \frac{\pi_{21}}{\pi_{11}+\pi_{21}}}{\frac{\pi_{12}}{\pi_{12}+\pi_{22}} \Big/ \frac{\pi_{22}}{\pi_{12}+\pi_{22}}}$$

$$= \frac{\omega_1/(1-\omega_1)}{\omega_2/(1-\omega_2)}$$

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, $\widehat{\text{RD}}$ 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

$$\widehat{\mathrm{Var}}(\log(\widehat{\mathrm{RR}})) = \frac{n_{12}}{n_{11}n_1} + \frac{n_{22}}{n_{21}n_2}$$

and

$$\log(\widehat{\mathrm{RR}}) \sim N(\log(\mathrm{RR}), \mathrm{Var}(\log(\mathrm{RR})))$$

- Therefore a $100(1-\alpha)\%$ CI for $\log(\mathrm{RR})$ 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

$$\mathrm{CI}_{\mathrm{lower}} = \frac{p_1}{p_2} \exp\left\{ -z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11} n_1} + \frac{n_{22}}{n_{21} n_2}} \right\}$$

$$\mathrm{CI}_{\mathrm{upper}} = \frac{p_1}{p_2} \exp\left\{ z_{1-\alpha/2} \sqrt{\frac{n_{12}}{n_{11} n_1} + \frac{n_{22}}{n_{21} n_2}} \right\}$$

- 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

$$\begin{aligned}\widehat{\mathrm{RD}} &= p_1 - p_2 \\ &= 36/190 - 252/3602 \\ &= 0.189 - 0.070 \\ &= 0.12\end{aligned}$$

  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

$$\widehat{\mathrm{RR}} = 0.189/0.070 = 2.7$$

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

              Frequency|yes     |no      |  Total
              ---------+--------+--------+
              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

$$\widehat{\mathrm{Var}}(\log(\widehat{\mathrm{OR}})) = \frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}$$

  and

$$\log(\widehat{\mathrm{OR}}) \sim N(\log(\mathrm{OR}), \mathrm{Var}(\log(\mathrm{OR})))$$

  (Woolf, 1955)

- Thus for large $n$, a $100(1-\alpha)\%$ CI is

$$\widehat{\mathrm{OR}} \exp\left\{ \pm z_{1-\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{21}} + \frac{1}{n_{12}} + \frac{1}{n_{22}}} \right\}$$

# 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

$$\widehat{\mathrm{OR}} = \frac{29 \times 1607}{205 \times 135} = 1.68$$

- 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 ouput:

```
                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
-----------------------------------------------------------------
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

$$H_0 : \pi_1 = \pi_2 \;\Leftrightarrow\; H_0 : \mathrm{OR} = 1$$

- $\widehat{\mathrm{OR}} = 1.71$; $X^2 = 5.34$ $(p = 0.02)$

- However, men work outdoors more than women

# Confounding: Example cont.

- Stratified analysis
- Outdoor occupation $\widehat{\text{OR}} = 1.06$

| | Malaria | No malaria | |
|---|---|---|---|
| Males | 53 | 15 | 68 |
| Females | 10 | 3 | 13 |
| | 63 | 18 | 81 |

- Indoor occupation $\widehat{\text{OR}} = 1.00$

| | 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 $2 \times 2$ tables.
- For each stratum, $j = 1, 2, \ldots, S$, 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

$$\mathrm{Var}(n_{11j}) = \frac{n_{1j}n_{2j}m_{1j}m_{2j}}{N_j^2(N_j - 1)}$$

- Let

$$O_j = n_{11j}; \;\; E_j = E(n_{11j}); \;\; V_j = \mathrm{Var}(n_{11j})$$

and

$$O = \sum_{j=1}^{S} O_j; \;\; E = \sum_{j=1}^{S} E_j; \;\; V = \sum_{j=1}^{S} V_j;$$

# Confounding: Mantel-Haenszel

- The Mantel-Haenszel statistic is given by

$$X^2_{\text{MH}} = \frac{(|O - E| - 0.5)^2}{V}$$

- Under $H_0 : \text{OR} = 1$ within strata, $X^2_{\text{MH}} \sim \chi^2_1$

$$C_\alpha = \{X^2_{\text{MH}} : X^2_{\text{MH}} > \chi^2_{1,1-\alpha}\}$$

- $X_{\text{MH}}$ 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

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} = \frac{\displaystyle\sum_{j=1}^{S} n_{11j}\, n_{22j}/N_j}{\displaystyle\sum_{j=1}^{S} n_{12j}\, n_{21j}/N_j}$$

# Confounding: Mantel-Haenszel

- Let

$$P_j = (n_{11j} + n_{22j})/N_j; \;\; Q_j = (n_{12j} + n_{21j})/N_j$$
$$R_j = (n_{11j}\, n_{22j})/N_j; \;\; W_j = (n_{12j}\, n_{21j})/N_j$$

- Then $\mathrm{Var}(\log(\widehat{\mathrm{OR}}_{\mathrm{MH}}))$ 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

$$\widehat{\mathrm{OR}}_{\mathrm{MH}} \exp\left\{ \pm z_{1-\alpha/2} \sqrt{\mathrm{Var}(\log(\widehat{\mathrm{OR}}_{\mathrm{MH}}))} \right\}$$

- Robins, Breslow, Greenland (Biometrics, 1986); See Rosner 1995 p 410

# Confounding: Malaria Example Revisited

- Unstratified: $X^2 = 5.34$
- Outdoor $\widehat{\text{OR}} = 1.06$; indoor $\widehat{\text{OR}} = 1.00$
- Outdoor:

$$O_1 = 53; \;\; E_1 = \frac{68 \times 63}{81} = 52.889;$$

$$V_1 = \frac{68 \times 13 \times 63 \times 18}{81^2 \times 80} = 1.9099$$

- Indoor:

$$O_2 = 35; \;\; E_2 = 34.9589; \;\; V_2 = 12.6620$$

# Confounding: Malaria Example cont.

- MH test statistic

$$X^2_{\text{MH}} = \frac{(|(53+35)-(52.889+34.9589)|-0.5)^2}{1.9099+12.6620}$$

$$= 0.008$$

without continuity correction $X^2_{\text{MH}} = 0.0016$

# 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, 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; $i = 1, 2, \ldots, n$; $j = 1, 2$

| | | $D^+$ | | |
|---|---|---|---|---|
| | | $Y_{i1} = 1$ | $Y_{i1} = 0$ | |
| $D^-$ | $Y_{i2} = 1$ | $n_{11}$ | $n_{12}$ | |
| | $Y_{i2} = 0$ | $n_{21}$ | $n_{22}$ | |
| | | | | $n$ |

# 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

$$\begin{aligned} \pi_1 - \pi_2 &= \Pr[Y_{i1} = 1] - \Pr[Y_{i2} = 1] \\ &= \Pr[Y_{i1} = 1, Y_{i2} = 0] - \Pr[Y_{i1} = 0, Y_{i2} = 1] \end{aligned}$$

# 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

$$\mathrm{M} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

- Under $H_0$, $\mathrm{M} \sim \chi^2_1$ if $n_{12} + n_{21}$ is sufficiently large (i.e. $\geq 30$)

$$C_\alpha = \{\mathrm{M} : \mathrm{M} > \chi^2_{1,1-\alpha}\}$$

$$p = \Pr[\chi^2_1 \geq m]$$

## 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

$$\mathrm{M} = \frac{(42-65)^2}{42+65} = 4.94$$

- 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 $\chi^2$ 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 $\pi_i = \Pr(\text{itching relief using lotion } i)$

$$H_0 : \pi_1 = \pi_2 \quad \text{vs.} \quad H_A : \pi_1 \neq \pi_2$$

- Exact p-value

$$p = 2 \sum_{k=0}^{6} \binom{16}{k} 2^{-16} = 2 \times 0.2272 = 0.4544$$

- Do not reject $H_0$

$$\mathrm{M} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(6 - 10)^2}{6 + 10} = 1$$

- R: mcnemar.test()

# Matched/Paired Observations: Example II cont.

```
proc freq order=data;
   tables lotion2*lotion1 / norow nocol nopercent;
   exact agree; weight count;
```

```
                    The FREQ Procedure

                Table of lotion2 by lotion1

            lotion2     lotion1

            Frequency|relief  |norelief|  Total
            ---------+--------+--------+
            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: $\mathrm{M} = 0$; $p = 1$

- We'll look at a measure of agreement (kappa statistic) later in the semester

# Matched or Paired Observations

- Odds ratio for matched data

$$\widehat{\mathrm{OR}}_{\mathrm{M}} = n_{21}/n_{12}$$

this is just $\widehat{\mathrm{OR}}_{\mathrm{MH}}$ with a stratum for each matched pair

- Confidence interval obtained by starting on the log scale

$$\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})) \approx \frac{1}{n_{12}} + \frac{1}{n_{21}}$$

- For $n_{12} + n_{21} \geq 30$, an approximate $100(1-\alpha)\%$ CI

$$\exp\left(\ln\left(\widehat{\mathrm{OR}}_{\mathrm{M}}\right) \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})}\right)$$

# Matched or Paired Observations: CMV Example

- Odds ratio estimate

$$\widehat{\mathrm{OR}}_{\mathrm{M}} = 65/42 = 1.55$$

- Corresponding estimate of variance of $\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})$

$$\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_{\mathrm{M}})) = \frac{1}{65} + \frac{1}{42} = 0.0392$$

- 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 | $\cdots$ | $c$ |
| | 1 | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1c}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2c}$ |
| $i$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rc}$ |

- Notation:

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \qquad n_{.j} = \sum_{i=1}^{r} n_{ij}$$

# 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

| | Practice | | | | |
|---|---|---|---|---|---|
| Reared | <5K | 5-49K | 50-99K | ≥100K | Total |
| <5K | 40 | 38 | 32 | 37 | 147 |
| 5-49K | 26 | 42 | 35 | 33 | 136 |
| 50-99K | 24 | 26 | 34 | 31 | 115 |
| ≥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 | | | | | |
|---|---|---|---|---|---|---|
| | $<20$ | 20-24 | 25-29 | 30-34 | $\geq 35$ | Total |
| Case | 320 | 1206 | 1011 | 463 | 220 | 3220 |
| Control | 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

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$$

- Test of independence, $X \perp Y$

$$\Pr[X = i, Y = j] = \Pr[X = i] \Pr[Y = j]$$

for $i = 1, \ldots, r;\ j = 1, \ldots, c$

# Contingency Tables

- For the breast cancer example, $H_0$ : distribution of age at first childbirth is the same for cases and controls

$$H_0 : \pi_{ij} = \pi_{i'j}; \;\; j = 1, 2, \ldots, c$$

- 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

$$E_{11} = n_{1\cdot} \frac{n_{\cdot 1}}{N} = \frac{n_{1\cdot} n_{\cdot 1}}{N}$$

    cases to be <20 years old

# Test of Independence or Association

- Under $H_0$, the expected frequency in the $(i, j)$ cell is

$$E_{ij} = \frac{n_{i.}n_{.j}}{N}$$

- Let

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

that is,

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - n_{i.}n_{.j}/N)^2}{n_{i.}n_{.j}/N}$$

# 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

| | Practice | | | | |
|---|---|---|---|---|---|
| Reared | <5K | 5-49K | 50-99K | ≥100K | Total |
| <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 |
| ≥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 | | | | | |
|---|---|---|---|---|---|---|
| | <20 | 20-24 | 25-29 | 30-34 | ≥35 | Total |
| Case | $\pi_{11}$ | $\pi_{12}$ | $\pi_{13}$ | $\pi_{14}$ | $\pi_{15}$ | 1 |
| Control | $\pi_{21}$ | $\pi_{22}$ | $\pi_{23}$ | $\pi_{24}$ | $\pi_{25}$ | 1 |

- Null hypothesis

$$H_0 : \pi_{1j} = \pi_{2j} \quad \text{for} \quad j = 1, 2, 3, 4, 5$$

- We can use the same statistic

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(c-1)}$$

# Breast Cancer Example cont.

- Expected frequencies

| | Age at first childbirth | | | | | |
|---|---|---|---|---|---|---|
| | $<20$ | 20-24 | 25-29 | 30-34 | $\geq 35$ | Total |
| Case | 416.6 | 1348.3 | 933.6 | 371.9 | 149.7 | 3220 |
| Control | 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 $\chi^2$ 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

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/N)^2}{n_{i\cdot}n_{\cdot j}/N}$$

$$= \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{\cdot 1}n_{2\cdot}n_{\cdot 2}}$$

- This is the *Pearson chi-square statistic* we saw for comparing two proportions

# Test for Trend

- Consider a $2 \times c$ table
- The $\chi^2$ 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) | | | |
|---|---|---|---|---|
| Culture | 1 | 2 | 3 | 4+ |
| Positive | 1 | 5 | 5 | 14 |
| Negative | 46 | 64 | 39 | 76 |
| Total | 47 | 69 | 44 | 90 |

- Let $\rho_j$ denote the conditional probability of being in row 1 given in column $j$
- For this catheter example, $\rho_j$ is the probability of a positive culture given in the $j^{\text{th}}$ 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:

$$x_j : j = 1, 2, \ldots, c$$

- Example: $x_j = j$
- In the breast cancer example, use the mid-range of age categories

# Test for Trend

- Let

$$[n_1 x] \equiv \sum_{j=1}^{c} n_{1j} x_j - \frac{n_{1\cdot} \sum_{j=1}^{c} n_{\cdot j} x_j}{N}$$

$$[x^2] \equiv \sum_{j=1}^{c} n_{\cdot j} x_j^2 - \frac{(\sum_{j=1}^{c} n_{\cdot j} x_j)^2}{N}$$

and

$$p \equiv \frac{n_{1\cdot}}{N}$$

- Then the chi-square test for trend (text, p 215) is

$$X^2_{\text{trend}} \equiv \frac{[n_1 x]^2}{[x^2] p(1-p)}$$

# Test for Trend

- Huh?
- Compute the average score in row 1:

$$\bar{x} \equiv \sum_{j=1}^{c} \frac{n_{1j} x_j}{n_{1.}}$$

- Compute the finite-sample expected value under the null:

$$E(\bar{x}) \equiv E(x) \equiv \sum_{j=1}^{c} \frac{n_{.j} x_j}{N}$$

# Test for Trend

- Compute the finite-sample variance:

$$\mathrm{Var}(\bar{x}) \equiv \left(\frac{1-f}{n_{1\cdot}}\right)\left[E(x^2) - E(x)^2\right]$$

where $f = n_{1\cdot}/N$ 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

$$X^2_{\text{trend}} = \frac{(\bar{x} - E(\bar{x}))^2}{\mathrm{Var}(\bar{x})}$$

# Test for Trend

- Under $H_0$,

$$X^2_{\text{trend}} \sim \chi^2_1$$

$$C_\alpha = \{X^2 : X^2 > \chi^2_{1,1-\alpha}\}$$

$$p = \Pr[\chi^2_1 > x^2]$$

- Note that here $\chi^2$ has 1 degree of freedom regardless of $c$

# Test for Trend: Example

- Catheter example

| | Duration | | | |
|---|---|---|---|---|
| Culture | 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

Frequency|
Col Pct  |       1|       2|       3|       4|  Total
---------+--------+--------+--------+--------+
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
```

# $\chi^2$ Test of Goodness of Fit

- Goal: Assess how well a particular model fits the data
- General form

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{\text{df}}$$

- $E_i$ computed under $H_0$
- Rejecting the null implies the model does not provide an adequate fit to the data

# $\chi^2$ 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 $\pi_1, \ldots, \pi_K$
- Let $n_i$ be the number of trials having outcome $i$, $i = 1, \ldots, K$, such that

$$n = \sum_{i=1}^{K} n_i$$

- Then

$$E(n_i) = n\pi_i \quad \text{and} \quad \text{Var}(n_i) = n\pi_i(1 - \pi_i)$$

# $\chi^2$ GOF: Genetics Example

- Example: Mendelian genetics hypothesizes a particular genotype should occur in the proportions

$$1 : 2 : 1 \text{ (dominant, heterozygous, recessive)}$$

- $H_0 : \pi_1 = 0.25,\ \pi_2 = 0.5,\ \pi_3 = 0.25$
- $H_A$ : 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 $n_1 = 21,\ \ n_2 = 62,\ \ n_3 = 17$

# $\chi^2$ 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 $H_0$,

$$E_1 = 0.25 \times 100 = 25$$

$$E_2 = 50$$

$$E_3 = 25$$

# $\chi^2$ 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 $H_0$

# $\chi^2$ GOF: Genetics Example Using SAS

```
proc freq order=data;
   tables genotype / testp=(25 50 25);
   weight count;
```

```
The FREQ Procedure

                                                Test     Cumulative    Cumulative
genotype        Frequency     Percent     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
```

# $\chi^2$ 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 |
| $\geq 110$ | 251 |
| Total | 14736 |

# $\chi^2$ 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

$$14736\,\Phi((50 - 80.7)/12) = 14736\,\Phi[-2.56] = 77.5$$

# $\chi^2$ 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 |
| $\geq 110$ | 251 | $\infty$ | 1 | 0.0073 | 107.7 |
| Total | 14,736 | | | | 14,736.0 |

# $\chi^2$ GOF: DBP Example cont.

- $H_0$ : data are from a normal distribution
- $H_A$ : data are not from a normal distribution
- Rejection region

$$C_\alpha = \{X^2 : X^2 \geq \chi^2_{K-S-1,1-\alpha}\}$$

where $S$ = number of parameters estimated

- For the DBP example, $K = 8$ and $S = 2$, so that

$$C_{0.05} = \{X^2 : X^2 \geq \chi^2_{5,0.95} = 11.07\}$$

# $\chi^2$ GOF: DBP Example cont.

- GOF test statistic

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{(57 - 77.5)^2}{77.5} + \cdots = 348.3$$

- Reject $H_0$

- Read Section 6.6.4 of the text about the distribution of $X^2$ when parameter estimation is necessary

# $\chi^2$ 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 $X_1^2, \ldots X_n^2$ are independent $\chi^2$ with $m_1, \ldots, m_n$ 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

| | Measured | | | | |
|---|---|---|---|---|---|
| Self-report | Under | Normal | Over | Obese | Total |
| 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 $\chi^2$ 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

$$\kappa = \frac{p_a - p_c}{1 - p_c}$$

- $\kappa$ 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 $\kappa$; 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 $\kappa$ needs to be context-dependent.

# Measure of Agreement: Kappa

- Under $H_0 : \kappa = 0$,

$$\mathrm{Var}(\kappa) = \frac{p_c + p_c^2 - N^{-3} \sum_{i=1}^{r} (n_{i\cdot}^2 n_{\cdot i} + n_{i\cdot} n_{\cdot i}^2)}{N(1 - p_c)^2}$$

- For moderate sample sizes,

$$z = \frac{\kappa}{\sqrt{\mathrm{Var}(\kappa)}} \sim N(0, 1)$$

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 $\kappa$

$$\mathrm{Var}(\kappa) = \frac{0.34 + 0.34^2 - 0.2631}{7455(0.66)^2} = 5.901 \times 10^{-5}$$

- 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
  - $\chi^2$
  - 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

- $H_0$: $Y_1, \ldots, Y_n \sim F_0(y)$
- 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

$$D = \max\{D_1, \ldots, D_n\}$$

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

$$F_0(y) = \begin{cases} 0 & \text{if} \quad y < 0 \\ y & \text{if} \quad 0 \le y \le 1 \\ 1 & \text{if} \quad 1 < y \end{cases}$$

- $n = 10$
- $C_{0.05} = \{D : D > 0.409\}$
- 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 $\mu$ and $\sigma^2$ 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 $\chi^2$ Goodness-of-Fit Tests

- If data are continuous, KS preferred. Why?
  - If sample size small, KS is exact, whereas $\chi^2$ relies on large sample approximation
  - KS test is more powerful than $\chi^2$ in most situations (Conover, *Practical Nonparametric Statistics*, 1980 p. 346)
  - Do not need to bin
- If discrete/categorical, $\chi^2$ 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) $\sigma^2$

- 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, $\lambda$
- If $Y \sim \text{Poisson}(\lambda)$, the probability mass function of $Y$ is

$$\Pr[Y = k] = \frac{e^{-\lambda}\lambda^k}{k!}$$

- $Y \in \{0, 1, 2, \ldots\}$
- The parameter $\lambda$ 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 $\pi$ 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 |
|---|---|---|---|---|---|
| | $N = 10$ | $N = 20$ | $N = 40$ | $N = 1000$ | PMF |
| $k$ | $\pi = 0.20$ | $\pi = 0.10$ | $\pi = 0.05$ | $\pi = 0.002$ | $\lambda = 2$ |
| 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 |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |

# The Poisson and Binomial Distributions

- Sketch of proof: Suppose on average $\mu$ 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 $\mu/N$

# The Poisson and Binomial Distributions

- Thus the probability of observing exactly $x$ events in the $N$ subintervals is

$$\frac{N(N-1)\cdots(N-x+1)}{x!}\left(\frac{\mu}{N}\right)^x\left(1-\frac{\mu}{N}\right)^{N-x} \tag{1}$$

- As $N \to \infty$,

$$N(N-1)\cdots(N-x+1) \approx N^x$$

and

$$\left(1-\frac{\mu}{N}\right)^{N-x} \approx \left(1-\frac{\mu}{N}\right)^{N} \to e^{-\mu}$$

- 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 $\lambda$ is

$$\left[\frac{1}{2}\chi^2_{2y;\,\alpha/2},\ \frac{1}{2}\chi^2_{2(y+1);\,1-\alpha/2}\right]$$

# Normal Approximations

- If $Y \sim \text{Poisson}(\lambda)$ and $\lambda$ large (say $\geq 100$), then

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

- Thus an approximate $100(1-\alpha)\%$ CI for $\lambda$ 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 $Y_1, Y_2, \ldots, Y_N$ iid Poisson($\lambda$), then

$$\sum_{i=1}^{N} Y_i \sim \text{Poisson}(N\lambda)$$

- Estimator for $\lambda$

$$\hat{\lambda} = \frac{1}{N} \sum_i Y_i$$

- If $(L, U)$ is a $100(1-\alpha)\%$ CI for $N\lambda$, then $(L/N, U/N)$ is a $100(1-\alpha)\%$ CI for $\lambda$. 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 $7\lambda$

$$\left[\frac{1}{2}\,\chi^2_{2\times 426;\,0.025},\ \frac{1}{2}\,\chi^2_{2\times 427;\,0.975}\right] = [386.50,\ 468.44]$$

- 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 $\lambda$

# 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 $\lambda$ is

$$\left[\left(\sqrt{Y} - 1\right)^2, \left(\sqrt{Y} + 1\right)^2\right]$$

- If we observe $y = 0$, a two-sided 90% CI for $\lambda$ is

$$[0, \tfrac{1}{2}\chi^2_{2;\,0.95}] \approx [0, 3.00]$$

- Thus if we observed $0$ events out of $N$ trials, the approximate upper bound on a two-sided 90% CI for $\lambda$ 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 $\lambda$s is not satisfied
- Want to test

$$H_0 : X_1, X_2, \ldots, X_k \sim \text{ Poisson } (\lambda)$$

# Homogeneity Test

- Construct a $\chi^2$ goodness of fit test using the following result

- Suppose $X_i \sim \text{Poisson}(\lambda_i)$ for $i = 1, 2, \ldots, k$
- Then the conditional distribution of $(X_1, \ldots, X_k)$ given $\sum_i X_i = N$ is multinomial with cell probabilities

$$\frac{\lambda_i}{\lambda_1 + \cdots + \lambda_k} \text{ for } i = 1, 2, \ldots, k$$

# Homogeneity Test

- Under $H_0$,

$$H_0 : X_1, X_2, \ldots, X_k \sim \text{ Poisson } (\lambda)$$

  the test statistic

$$T = \frac{\sum_i (X_i - \bar{X})^2}{\bar{X}} \sim \chi^2_{k-1}$$

- Equivalent form

$$T = \frac{(k-1)s^2}{\bar{X}}$$

- *Poisson homogeneity/heterogeneity/dispersion test*

# Homogeneity Test

- The $\chi^2$ approximation improves as $\lambda$, $k$ 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 $\Pr[\chi^2_6 > 5.93] = 0.43$, fail to reject $H_0$

# Negative-Binomial Distribution

- Over-dispersion may be due to heterogeneity of $\lambda$s
- That is, $\lambda$ is no longer a constant, but a random variable
- If $\lambda$ 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

$$Y = \alpha + \beta X$$

- $\alpha$ = intercept; value of $Y$ when $X = 0$
- $\beta$ = 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

$$Y = \alpha + \beta X + \epsilon$$

$$\epsilon = Y - \alpha - \beta X$$

- $\epsilon$ is the vertical distance from $Y$ to the line defined by $\alpha + \beta X$

# Simple Linear Model with Error

# Model Assumptions

- Data are $(Y_i, X_i)$; $i = 1, 2, \ldots, N$
- Assumptions:

  1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
  2. $X$s are fixed constants
  3. $\epsilon_i$ iid $N(0, \sigma^2)$

# Least Squares Estimation

- Least squares estimators are values of $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$
- Can also derive these estimators via maximum likelihood

# Least Squares Estimation

- For $\alpha$:

$$\frac{\partial \sum_i \epsilon_i^2}{\partial \alpha} = -2 \sum_i (Y_i - \alpha - \beta X_i)$$

$$= -2N\bar{Y} + 2N\alpha + 2N\beta\bar{X}$$

- For $\beta$:

$$\frac{\partial \sum_i \epsilon_i^2}{\partial \beta} = -2 \sum_i (Y_i - \alpha - \beta X_i) X_i$$

$$= -2 \sum_i X_i Y_i + 2\alpha \sum_i X_i + 2\beta \sum_i X_i^2$$

# Least Squares Estimation

- Two equations with two unknowns

$$-2N\bar{Y} + 2N\alpha + 2N\beta\bar{X} = 0 \tag{1}$$

$$-2\sum_i X_i Y_i + 2\alpha \sum_i X_i + 2\beta \sum_i X_i^2 = 0 \tag{2}$$

- 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

$$\hat{\beta}(\sum_i X_i^2 - N\bar{X}^2) = \sum_i X_i Y_i - N\bar{X}\bar{Y}.$$

- 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)

$$\hat{\sigma}^2 = s^2_{y \cdot x} = \frac{1}{N-2} \sum_i (Y_i - \hat{Y}_i)^2$$

$$= \frac{1}{N-2} \sum_i r_i^2$$

# Example: SBP and Age

$$\bar{Y} = 132.4; \quad \bar{X} = 44.8$$

$$\sum_i X_i Y_i = 54461; \quad \sum_i X_i^2 = 20463$$

$$\hat{\beta} = \frac{54461 - 9(132.4)(44.8)}{20463 - 9(44.8)^2} = 0.45$$

$$\hat{\alpha} = 132.4 - 0.45(44.8) = 112.3$$

$$\hat{Y}_i = 112.3 + 0.45 X_i$$

$$s^2_{y \cdot x} = 3.21$$

# 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.

$$\hat{\boldsymbol{\beta}} = \mathbf{0}$$

# $\hat{\beta} > 0$

# $\hat{\beta} < 0$

$$\hat{\boldsymbol{\beta}} = \mathbf{0}$$

# 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 $\beta$

$$\hat{\beta} \pm z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}$$

- 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 $\sigma^2$ is unknown, use $s^2_{y\cdot x}$ and $t_{N-2}$

- $100(1-\alpha)\%$ CI for $\beta$

$$\hat{\beta} \pm t_{N-2,1-\alpha/2}\sqrt{s^2_{y\cdot x}/\sum_i (X_i - \bar{X})^2}$$

- Test for $H_0 : \beta = \beta_0$

$$t = \frac{\hat{\beta} - \beta_0}{\sqrt{s^2_{y\cdot x}/\sum_i (X_i - \bar{X})^2}}$$

## 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

                                   Sum of           Mean
Source                    DF      Squares         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

                     Parameter       Standard
Variable     DF       Estimate          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 $\mu_x$:

$$\begin{aligned}\hat{\mu}_x &= \hat{\alpha} + \hat{\beta}x \\ &= \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta}x \\ &= \bar{Y} + \hat{\beta}(x - \bar{X})\end{aligned}$$

- $E(\hat{\mu}_x) = \mu_x$

# Confidence Interval for $E(Y|X = x)$

- Recall that $\bar{Y}$ and $\hat{\beta}$ are independent normally distributed random vaiables
- Thus $\hat{\mu}_x$ is normally distributed and

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_x) &= \mathrm{Var}(\bar{Y}) + (x - \bar{X})^2 \,\mathrm{Var}(\hat{\beta}) \\
&= \frac{\sigma^2}{N} + \frac{\sigma^2 (x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \\
&= \sigma^2 \left[ \frac{1}{N} + \frac{(x - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right]
\end{aligned}$$

## Confidence Interval for $E(Y|X = x)$

- Therefore, a $100(1-\alpha)\%$ CI for $\mu_x$ is

$$\hat{\mu}_x \pm t_{N-2,1-\alpha/2}\sqrt{s^2_{y\cdot x}\left\{\frac{1}{N} + \frac{(x-\bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right\}}$$

- Note that $\mathrm{Var}(\hat{\mu}_x)$ 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., $4^{\text{th}}$ 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 $Y_x - \hat{Y}_x$

$$E(Y_x - \hat{Y}_x) = \alpha + \beta x - (\alpha + \beta x) = 0$$

$$\mathrm{Var}(Y_x - \hat{Y}_x) = \mathrm{Var}(Y_x) + \mathrm{Var}(\hat{Y}_x) - 2\mathrm{Cov}(Y_x, \hat{Y}_x)$$

# Prediction

- Because $Y_x$ is not part of the sample, $Y_x$ and $\hat{Y}_x$ are independent
- Therefore

$$\begin{aligned}\mathrm{Var}(Y_x - \hat{Y}_x) &= \sigma^2 + \sigma^2 \left( \frac{1}{N} + \frac{(x-\bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right) \\ &= \sigma^2 \left( 1 + \frac{1}{N} + \frac{(x-\bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right)\end{aligned}$$

# Prediction

- Because $\epsilon$ 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 $\sigma^2$ 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
- $r^2 = 1 - \text{SSE/SST}$ typically increases with range/spacing of $X$

# Examples of $r^2$

# Examples of $r^2$: $Y = 0 + 1 \cdot X + \epsilon$, $\epsilon \sim N(0, 1)$

# Examples of $r^2$: $Y = 0 + 1 \cdot X + \epsilon$, $\epsilon \sim N(0, 4)$

# 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 $H_0 : \beta = 0$,

$$t = \frac{\hat{\beta}}{\sqrt{s^2_{y.x} / \sum_i (X_i - \bar{X})^2}} \sim t_{N-2}$$

- Equivalently,

$$t = \frac{[XY]/[X^2]}{\sqrt{s^2_{y.x}/[X^2]}} \sim t_{N-2}$$

- In general, if $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$. Thus

$$t^2 = \frac{[XY]^2/[X^2]}{s^2_{y.x}} \sim F_{1,N-2}$$

# Analysis of Variance

- Note

$$\text{SSR} = \sum(\hat{Y}_i - \overline{Y})^2 = \sum(\hat{\alpha} + \hat{\beta}X_i - \overline{Y})^2$$

$$= \sum(\overline{Y} - \hat{\beta}\overline{X} + \hat{\beta}X_i - \overline{Y})^2$$

$$= \sum \hat{\beta}^2(X_i - \overline{X})^2$$

$$= \frac{[XY]^2}{[X^2]^2}\sum(X_i - \overline{X})^2 = \frac{[XY]^2}{[X^2]}$$

- Thus

$$t^2 = \frac{\text{SSR}}{\text{MSE}} = \frac{\text{SSR}}{\text{SSE}/(N-2)}$$

# Analysis of Variance

- If $\beta = 0$ then

$$\frac{\text{SSR}}{\sigma^2} \sim \chi^2_1 \quad \perp \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2_{N-2}$$

(*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

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

- Linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

# 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:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$$

# Matrix Formulation

- Therefore

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- We can also show

$$\text{SST} = \boldsymbol{Y}'\boldsymbol{Y} - \frac{1}{N}\boldsymbol{Y}'\boldsymbol{J}\boldsymbol{Y}$$

$$\text{SSR} = \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - \frac{1}{N}\boldsymbol{Y}'\boldsymbol{J}\boldsymbol{Y}$$

$$\text{SSE} = \boldsymbol{Y}'\boldsymbol{Y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$$

where $\boldsymbol{J}$ 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 $i = 1, \ldots, n_1$ and $Y_{2i}$ for $i = 1, \ldots, n_2$

- 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

$$N = n_1 + n_2$$

$$(Y_1, \ldots, Y_{n_1}) = (Y_{11}, \ldots, Y_{1n_1})$$

$$(Y_{n_1+1}, \ldots, Y_N) = (Y_{21}, \ldots, Y_{2n_2})$$

$$X_i = \begin{cases} 1 & \text{if in group 1} \\ \\ 0 & \text{if in group 2} \end{cases}$$

# Linear Regression and Two Sample t-test

- Consider the regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i; \quad i = 1, 2, 3, \ldots, N$$

- 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

$$\hat{\beta} = \frac{(1-\bar{X})\sum_{i=1}^{n_1} Y_i}{[X^2]} + \frac{(-\bar{X})\sum_{i=n_1+1}^{N} Y_i}{[X^2]}$$

$$= \bar{Y}_1 - \bar{Y}_2$$

- We can show that

$$s^2_{y \cdot x} = s^2_p$$

# Linear Regression and Two Sample t-test

- Therefore:

$$t = \frac{\hat{\beta}}{\sqrt{s^2_{y\cdot x} / \sum_i (X_i - \bar{X})^2}}$$

$$= \frac{\bar{Y}_1 - \bar{Y}_2}{s_p \sqrt{N/(n_1 n_2)}}$$

# 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. $\epsilon_i$ 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 $\epsilon_i$ 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: $FEV_1$

# 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

$$Y = \alpha + \beta_{y \cdot x} X + \epsilon$$

$$X \perp \epsilon; \;\; \mathrm{Var}(X) = \delta^2$$

- 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

$$\mathrm{Cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$$

# Regression: $X$ Random

- Now

$$\beta_{y\cdot x} = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}$$

- Proof: We have $\mathrm{Cov}(a + bW, U) = b\,\mathrm{Cov}(W, U)$ and $\mathrm{Cov}(W, U + V) = \mathrm{Cov}(W, U) + \mathrm{Cov}(W, V)$
- Thus

$$\begin{aligned}
\mathrm{Cov}(Y, X) &= \mathrm{Cov}(\alpha + \beta_{y\cdot x} X + \epsilon, X) \\
&= \beta_{y\cdot x}\mathrm{Cov}(X, X) + \mathrm{Cov}(\epsilon, X) \\
&= \beta_{y\cdot x}\mathrm{Var}(X)
\end{aligned}$$

# Measurement Error

- Instead of observing $X$, we observe

$$W = X + U$$

where $U$ is a random variable with

$$E(U) = 0, \quad \mathrm{Var}(U) = \tau^2$$

$$U \perp X, \quad U \perp Y$$

- Then

$$\begin{aligned} \mathrm{Cov}(W, Y) &= \mathrm{Cov}(X + U, Y) \\ &= \mathrm{Cov}(X, Y) + \mathrm{Cov}(U, Y) = \mathrm{Cov}(X, Y) \end{aligned}$$

# Measurement Error

- By independence

$$\mathrm{Var}(W) = \mathrm{Var}(X) + \mathrm{Var}(U) = \delta^2 + \tau^2$$

- Thus

$$\begin{aligned}
\beta_{y\cdot w} &= \frac{\mathrm{Cov}(Y,\, W)}{\mathrm{Var}(W)} \\
&= \frac{\mathrm{Cov}(Y,\, X)}{\delta^2 + \tau^2} \\
&= \frac{\delta^2}{\delta^2 + \tau^2} \frac{\mathrm{Cov}(Y,\, X)}{\delta^2} \\
&= \frac{\delta^2}{\delta^2 + \tau^2} \beta_{y\cdot x}
\end{aligned}$$

# 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 $R_{\text{rel}}$ 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

$$\mathrm{Var}(\tilde{\beta}) = R_{\mathrm{rel}}^{-2}\,\mathrm{Var}(\hat{\beta}_{y\cdot w})$$

the $t$-statistic for testing $H_0 : \beta_{y\cdot x} = 0$ is

$$t_{y\cdot x} = \frac{\tilde{\beta}}{\sqrt{\mathrm{Var}(\tilde{\beta})}} = \frac{R_{\mathrm{rel}}^{-1}\hat{\beta}_{y\cdot w}}{\sqrt{R_{\mathrm{rel}}^{-2}\,\mathrm{Var}(\hat{\beta}_{y\cdot w})}} = t_{y\cdot w}$$

# Measurement Error

- Suppose there are $k$ independent measures of $W$ made on each person in a study
- It can be shown that

$$\mathrm{Var}(\bar{W}_k) = \delta^2 + \frac{\tau^2}{k}$$

- Therefore

$$\beta_{y \cdot \bar{w}_k} = \frac{\delta^2}{\delta^2 + \tau^2/k} \beta_{y \cdot x} \to \beta_{y \cdot x} \quad \text{as} \quad k \to \infty$$

# 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 $\epsilon$

# 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 $\tau$

# 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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i,$$

$$i = 1, 2, \ldots, N$$

- Data are $(Y_i, \boldsymbol{X}_i)$; $i = 1, 2, \ldots, N$, where $\boldsymbol{X}_i$ 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. $\epsilon_i$ iid $N(0, \sigma^2)$

# Multiple Linear Regression Model

Multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i,$$

$$i = 1, 2, \ldots, N$$

Interpretation of parameters:

- $\beta_j$ is the change in the expected value of $Y$ when the $j^{\text{th}}$ $X$ variable increases by one unit, with all the other $X$ variables being held contact
- If the $j^{\text{th}}$ $X$ 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 $j^{\text{th}}$ $X$ is 1 versus when it is 0

# Matrix Formulation

- Let

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N1} & X_{N2} & \dots & X_{Nk} \end{pmatrix},$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

# Matrix Formulation

- Linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- The least squares estimators are the solutions to the set of equations:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$$

- Therefore, as in the simple linear regression case:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

- 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

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

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 Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| 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 Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| 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

$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Note:

$$\rho = \frac{\beta_{y \cdot x}\sigma_X}{\sigma_Y} = \frac{\beta_{x \cdot y}\sigma_Y}{\sigma_X}$$

# Correlation

- Estimate $\rho$ 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

# 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

$$
\begin{aligned}
t &= \frac{\hat{\beta}_{y\cdot x}}{s_{y\cdot x}/\sqrt{[X^2]}} = \frac{[XY]/[X^2]}{s_{y\cdot x}/\sqrt{[X^2]}} \\
&= \frac{[XY]/\sqrt{[X^2]}}{s_{y\cdot x}} = \frac{r\sqrt{[Y^2]}}{s_{y\cdot x}} \\
&= \frac{r s_Y \sqrt{N-1}}{\sqrt{(1-r^2)s_Y^2(N-1)/(N-2)}} \\
&= \frac{r}{\sqrt{(1-r^2)/(N-2)}}
\end{aligned}
$$

# Correlation

- In general,

$$t = \frac{r}{\sqrt{(1-r^2)/(N-2)}} \sim t_{N-2} \tag{1}$$

  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 $\rho$

$$z_r = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \;\Rightarrow\; e^{2z_r} = \frac{1+r}{1-r} \;\Rightarrow\; r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$$

$$z_L = z_r - z_{1-\alpha/2}\sqrt{1/(N-3)}$$

$$z_U = z_r + z_{1-\alpha/2}\sqrt{1/(N-3)}$$

$$r_L = \frac{e^{2z_L}-1}{e^{2z_L}+1}; \quad r_U = \frac{e^{2z_U}-1}{e^{2z_U}+1}$$

# 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 H0:Rho=0 |
|---|---|---|---|---|
| x | y | 0.056350 | 0.425524 | 0.0119 |

# 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

  $\rho_1$ = population correlation for natural pairs

  $\rho_2$ = 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 $H_0$

$$H = T_2 - \frac{T_1^2}{\sum (n_i - 3)} \sim \chi^2_{k-1}$$

- 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 $\tau$
- 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

$$r_s = \frac{\sum(R_{1i}-\bar{R}_1)(R_{2i}-\bar{R}_2)}{\sqrt{\sum_i(R_{1i}-\bar{R}_1)^2 \sum_i(R_{2i}-\bar{R}_2)^2}}$$

$$= 1 - \frac{6\sum d_i^2}{N^3-N}$$

  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:

  $R_{1i} = R_{2i}$ for all $i$ $\Rightarrow$ $d_i = 0$ for all $i$ $\Rightarrow$ $r_s = 1$

# Spearman Correlation

- Suppose $N$ is odd and $N = 2m + 1$
- Then the most extreme discordant rankings are

| $i$ | $1$ | $2$ | $\cdots$ | | | | | | $N$ |
|---|---|---|---|---|---|---|---|---|---|
| $R_{1i}$ | $1$ | $2$ | $\cdots$ | $m$ | $m+1$ | $m+2$ | $\cdots$ | $2m$ | $2m+1$ |
| $R_{2i}$ | $2m+1$ | $2m$ | $\cdots$ | $m+2$ | $m+1$ | $m$ | $\cdots$ | $2$ | $1$ |
| $d_i$ | $-2m$ | $2-2m$ | $\cdots$ | $-2$ | $0$ | $2$ | $\cdots$ | $2m-2$ | $2m$ |

# Spearman Correlation

- Under this configuration

$$
\begin{aligned}
\sum_{i=1}^{N} d_i^2 &= 4m^2 + 4(m-1)^2 + \cdots + 4(1)^2 + 0 \\
&\qquad\qquad + 4(1)^2 + \cdots + 4(m-1)^2 + 4m^2 \\
&= 8\sum_{j=1}^{m} j^2 \\
&= 8m(m+1)(2m+1)/6 \\
&= \left(4 \times \tfrac{N-1}{2} \times \tfrac{N+1}{2} \times N\right)/3 \\
&= (N^3 - N)/3
\end{aligned}
$$

# 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

$$r_s = 1 - \frac{6(2^2 + (-3)^2 + \cdots + 1^2)}{10^3 - 10} = 1 - \frac{6(182)}{990} = -0.103$$

# 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

$$H_0 : X \perp Y \quad \text{vs.} \quad H_A : X \not\perp Y$$

  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 $\tau$

- Kendall's $\tau$: Another rank correlation statistic
- Data: $(X_i, Y_i)$ for $i = 1, 2, \ldots, N$
- 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 $\tau$

- 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 $\tau$

- 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 $\tau$ 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 $\tau$

# Kendall's $\tau$

- $H_0 : \tau = 0$ 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 $\tau$

- 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 $\tau$: 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 $\tau$

- $C_{0.05} = \{r_k : |r_k| \geq 0.64\}$
- 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 $\tau$

- If $N$ is sufficiently large ($\geq 10$), 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 $\tau$

- 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 $i^{\text{th}}$ set of ties for $z = x, y$

# Kendall's $\tau$

- 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* $\tau_b$

# Kendall's $\tau$: 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 $\tau$: 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       1.00000
                  0.0008
```

# Kendall's $\tau$: 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 $\tau$

- 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

$$Z = \frac{P - Q}{\sqrt{\mathrm{Var}(P - Q)}} \sim N(0, 1)$$

# Kendall's $\tau$: Example

- In general, $\mathrm{Var}(P-Q)$ equals

$$\frac{N(N-1)(2N+5)}{18} - \sum_i \frac{t_{xi}(t_{xi}-1)(2t_{xi}+5)}{18} - \cdots$$

- For tobacco example, $\mathrm{Var}(P-Q)$ 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 $\rho$,

$$r \xrightarrow{p} \rho \qquad r_s \xrightarrow{p} \frac{6}{\pi} \arcsin(\rho/2) \qquad r_k \xrightarrow{p} \frac{2}{\pi} \arcsin(\rho)$$

  (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 $j^{\text{th}}$ observation in the $i^{\text{th}}$ group
- $i = 1, \ldots, K; \;\; j = 1, \ldots, n_i$
- 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 $\sigma^2$ 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 $\bar{Y}_i$.s will estimate $\sigma^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2}{K - 1}$$

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 $H_0$:

$$(N-K)s_p^2/\sigma^2 \sim \chi^2_{N-K}$$

$$(K-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{K-1}$$

  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 $\mu$ 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 $i = 1, 2, \ldots, K;\ \ j = 1, 2, \ldots, n_i$ 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 $i = 1, 2, \ldots, K;\ \ j = 1, 2, \ldots, n_i$ where

$$\mu = \frac{1}{N} \sum_{i=1}^{K} n_i \mu_i \tag{1}$$

$$\alpha_i = \mu_i - \mu$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

- Note typo in the text on page 363
- Here $\alpha_i$ 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

$$\begin{aligned}
Y_{1j} &= \mu + 1\alpha_1 + \epsilon_{1j} \\
Y_{2j} &= \mu + 1\alpha_2 + \epsilon_{2j} \\
Y_{3j} &= \mu + 1\alpha_3 + \epsilon_{3j} \\
Y_{4j} &= \mu - \frac{n_1}{n_4}\alpha_1 - \frac{n_2}{n_4}\alpha_2 - \frac{n_3}{n_4}\alpha_3 + \epsilon_{4j}
\end{aligned}$$

# Model Equivalence

- Equivalence of null hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \;\Leftrightarrow\; H_0 : \alpha_i = 0; \;\; i = 1, 2, \ldots, K$$

- $\alpha_i$ is called the $i^{\text{th}}$ *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

$$\begin{aligned} Y_{1j} &= \mu_1 + \epsilon_{1j} \\ Y_{ij} &= \mu_1 + (\mu_i - \mu_1) + \epsilon_{ij}, \quad i = 2, 3, \ldots, K \\ &= \mu_1 + \beta_i + \epsilon_{ij}, \quad i = 2, 3, \ldots, K \end{aligned}$$

for

$$j = 1, 2, \ldots, n_i$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2) \text{ for all } i, j$$

- Null hypothesis:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$$

# 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,

$$\begin{aligned}\text{SST} &= \text{SSA} + \text{SSW} \\ &= (K-1)\hat{\sigma}^2 + (N-K)s_p^2\end{aligned}$$

- 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 $H_0 : \alpha_1 = \cdots = \alpha_K = 0$,

$$E\left(\sum_{i=1}^{K} n_i(\bar{Y}_{i\cdot} - \bar{Y})^2\right) = (K-1)\sigma^2$$

# 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: $E$(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 $\alpha$ 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 $\alpha$ level and each null hypothesis is true

| | $\alpha$ | | |
|---|---|---|---|
| $n$ | 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

$$t_{ij} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\sqrt{\mathrm{MSE}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

- 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 Between Means | Simultaneous 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 $Y_1, \ldots, Y_k$ iid $N(\mu, \sigma^2)$
- Let $s$ be an estimator for $\sigma$ with $m$ degrees of freedom, $s \perp Y_1, \ldots, Y_k$
- 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 Between Means | Simultaneous 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 $A_1, A_2, \ldots, A_n$ be a set of events
- Bonferroni inequality

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \Pr(A_i)$$

- Let $A_i$ be the event that we reject $H_{0i}$ when $H_{0i}$ is true for $i = 1, 2, \ldots, n$

$$\Pr(A_i) = \alpha_i$$

# Bonferroni Method

- Probability of at least one Type I error

$$\Pr(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \alpha_i$$

- If $\alpha_i = \alpha^*$ for all $i$,

$$\sum_{i=1}^{n} \alpha_i = n\alpha^*$$

- If we want $\Pr(A_1 \cup \cdots \cup A_n) \leq \alpha$, 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 $\mu_i$

$$\bar{Y}_{i\cdot} \pm t_{N-K;1-\alpha/2}\sqrt{\text{MSE}/n_i}$$

- 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

$$\frac{\hat{L} - L}{\sqrt{\mathrm{MSE} \sum_i c_i^2 / n_i}} \sim t_{N-K}$$

# Conclusion

- Factor level means, that is, $\mu_1, \mu_2, \ldots$: 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

$$F_{\text{MAX}} = \frac{s^2_{\max}}{s^2_{\min}}, \; C = \frac{s^2_{\max}}{\sum s_i^2}$$

    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

                          Sum of        Mean
        Source        DF     Squares      Square    F Value    Pr > F

        group          3      0.8003      0.2668       0.19    0.9001
        Error         19     26.3125      1.3849

              Level of           -------------age-------------
              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 $(\bar{y}_{i\cdot}, s_i)$, $(\bar{y}_{i\cdot}, s_i^2)$, $(\bar{y}_{i\cdot}^2, s_i)$; 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 $\lambda$

$$Y_\lambda = \begin{cases} k_1(Y^\lambda - 1) & \text{for } \lambda \neq 0 \\ k_2 \log(Y) & \text{for } \lambda = 0 \end{cases}$$

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 $\lambda$ 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 $i = 1, \ldots, K;\ j = 1, \ldots, n_i$.

- $\epsilon_{ij}$ are independent and identically distributed with mean zero, but not necessarily normal

# Kruskal-Wallis

- Same hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \quad \text{vs.} \quad H_A : \text{ at least one inequality}$$

- Pool all $N$ observations and rank from smallest to largest
- Let $R_{ij}$ be the rank of the $j^{\text{th}}$ obs in the $i^{\text{th}}$ group
- Let $\bar{R}_i = \sum_{j=1}^{n_i} R_{ij}/n_i$ equal the average rank in the $i^{\text{th}}$ group
- Let $\bar{R}$ denote the overall average rank. What must this equal?

# Kruskal-Wallis

- The Kruskal-Wallis test statistic is

$$T_{\mathrm{KW}} = \frac{12 \sum_{i=1}^{K} n_i (\bar{R}_i - \bar{R})^2}{N(N+1)}$$

- Equivalently

$$T_{\mathrm{KW}} = \frac{12 \sum_{i=1}^{K} (\sum_{j=1}^{n_i} R_{ij})^2 / n_i}{N(N+1)} - 3(N+1)$$

- Reject $H_0$ for large values of $T_{\mathrm{KW}}$

# Kruskal-Wallis

- Under $H_0$, if the $n_i$ are moderately large (rule of thumb: $n_i \geq 5$), then

$$T_{\mathrm{KW}} \sim \chi^2_{K-1}$$

- If the $n_i$ are small, the exact distribution of $T_{\mathrm{KW}}$ 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

| $R_{1j}$ | $R_{2j}$ | $R_{3j}$ | $\sum_i R_{i.}^2/n_i$ | $T_{\mathrm{KW}}$ |
|---|---|---|---|---|
| 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 |

| $k$ | $\Pr[T_{\mathrm{KW}} = k]$ |
|---|---|
| 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
- $T_{\text{KWadj}}$ will also be approximately $\chi^2_{K-1}$

# Kruskal-Wallis: Example

- A study was conducted to compared 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 $\mu_i$ denote the mean temperature change for dose $i$
- $H_0 : \mu_1 = \mu_2 = \mu_3$

## Kruskal-Wallis: Example cont.

- Distribution of $T_{\mathrm{KW}}$ (Owen 1962, page 422; Kruskal, Wallis 1952, JASA, Table 6.1)

| $k$ | $\Pr[T_{\mathrm{KW}} \geq k]$ |
|---|---|
| 4.50 | 0.102 |
| 4.56 | 0.100 |
| 5.66 | 0.051 |
| 5.78 | 0.049 |
| 7.98 | 0.010 |
| 8.00 | 0.009 |

- $C_{0.05} = \{T_{\mathrm{KW}} \geq 5.78\}$

# Kruskal-Wallis: Example cont.

| Low | | Med | | High | |
|---|---|---|---|---|---|
| $\Delta$Temp | R | $\Delta$Temp | R | $\Delta$Temp | 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 |

- $R_{1\cdot} = 63, \quad R_{2\cdot} = 32, \quad R_{3\cdot} = 25$

# Kruskal-Wallis: Example cont.

- Therefore

$$T_{\mathrm{KW}} = \frac{12(63^2/5 + 32^2/5 + 25^2/5)}{15(16)} - 3(16) = 8.18$$

- 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

                      Sum of     Expected      Std Dev         Mean
 dose        N       Scores     Under H0     Under H0        Score
 ----------------------------------------------------------------
 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

$$F_{\mathrm{R}} = \frac{(N-K)T_{\mathrm{KW}}}{(K-1)(N-1-T_{\mathrm{KW}})}$$

- 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

$j = 1, 2, \ldots, n_i$ denoting the $j^{\text{th}}$ patient in the $i^{\text{th}}$ group

$n_1 = 14, \ n_2 = 13, \ n_3 = 12$

$Y_{ij}$ survival time in days

$\mu_i$ 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 Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| degree | 2 | 452324 | 226162 | 2.46 | 0.0993 |
| Error | 36 | 3304203 | 91783.4 | | |

# QQ Plot

Normal Q−Q Plot
Sample Quantiles
−200
0
200
400
600
800
1000
−2
−1
0
1
2
Theoretical Quantiles

# Box-Cox Transformation

- Recall: Family of transformations indexed by $\lambda$

$$Y_\lambda = \begin{cases} k_1(Y^\lambda - 1) & \text{for } \lambda \neq 0 \\ k_2 \log(Y) & \text{for } \lambda = 0 \end{cases}$$

  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 $\lambda$ 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

                        Sum of        Mean
  Source        DF     Squares      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

# 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 ***.

                      Difference
       degree            Between     Simultaneous 90%
     Comparison            Means    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 ***.

                          Difference
          degree             Between     Simultaneous 90%
        Comparison             Means    Confidence Limits

      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

$$H_0 : (\mu_1 + \mu_2)/2 - \mu_3 = 0 \quad \text{vs.} \quad H_A : (\mu_1 + \mu_2)/2 - \mu_3 \neq 0$$

- 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;
```

| Parameter | Estimate | Standard 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 | 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 $\alpha$
  - The null hypothesis
  - The test statistic and its distribution
  - The value $\theta_A$ that we want to be able to detect
  - The desired power to detect this $\theta_A$
  - 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. $\alpha$
  - Obtain estimates from pilot studies or studies done elsewhere, e.g. $\theta_A$, variances, covariances
  - $\theta_A$ 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 $\theta_0$ is a known constant
- Usually,

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

- Once the data are collected, we compute a statistic related to $\theta$, 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
- $E(Y) = \mu, \;\; \mathrm{Var}(Y) = \sigma^2$

$$H_0 : \mu = \mu_0 \;\; \text{vs.} \;\; H_A : \mu > \mu_0$$

$$S(\hat{\theta}) = Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{N}}$$

# 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 $\mu_A$

# One Sample $Z$ Test

- Under $H_A : \mu = \mu_A$

$$Z' = \frac{\overline{Y} - \mu_A}{\sigma/\sqrt{N}} \sim N(0,1)$$

- Note

$$Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{N}} = Z' + \frac{\mu_A - \mu_0}{\sigma/\sqrt{N}}$$

- 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

$$1 - \beta = \Pr[Z > z_{1-\alpha} \mid \mu = \mu_A]$$

$$= \Pr[Z' > z_{1-\alpha} + \frac{\sqrt{N}(\mu_0 - \mu_A)}{\sigma} \mid \mu = \mu_A]$$

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

$$\Delta = \frac{|\mu_0 - \mu_A|}{\sigma}$$

- $\Delta$ is called the *standardized distance* or *difference*

# One Sample $Z$ Test

- To apply this formula, we need to know $\alpha$, $\beta$, $|\mu_0 - \mu_A|$, and $\sigma^2$
- 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 $\sigma^2$: Need data from the literature or a pilot study; note that we need the variance of the difference (after − before)
- The choice of $\mu_A$ 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 $\alpha$, $\beta$, $\sigma^2$, and $\mu_A$
- Note that $(1.96 + 1.28)^2 \approx 10.5$, so that

$$N \approx \frac{10.5}{\Delta^2}$$

# One Sample $Z$ Test

| $\alpha$ | $1-\beta$ | $\sigma^2$ | $\mu_A$ | $N$ | $\alpha$ | $1-\beta$ | $\sigma^2$ | $\mu_A$ | $N$ |
|---|---|---|---|---|---|---|---|---|---|
| 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, $\sigma$ 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 $\nu$ degrees of freedom and non-centrality parameter $\lambda$

# One Sample $t$ Test

- Consider $H_0 : \mu = 0$ versus $H_A : \mu = \mu_A$ for $\mu_A \neq 0$
- Under $H_A$,

$$\overline{Y} \sim N(\mu_A, \sigma^2/N)$$

- Thus

$$\frac{\overline{Y}\sqrt{N}}{\sigma} \sim N\left(\mu_A \frac{\sqrt{N}}{\sigma}, 1\right)$$

- Recall

$$\frac{(N-1)s^2}{\sigma^2} \sim \chi^2_{N-1}$$

# One Sample $t$ Test

- Because $\overline{Y} \perp s^2$,

$$T = \frac{\overline{Y}}{s/\sqrt{N}} \sim t_{N-1,\lambda}$$

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;
   onesamplefreq 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 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

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_A : \mu_1 \neq \mu_2$$

- Assume homogeneity of variance, $\sigma^2$ 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: $\boldsymbol{\sigma}$ Unknown

- What if the variance is not known?
- For $N_1 = N_2 = N$, one can show that

$$\frac{\overline{Y}_1 - \overline{Y}_2}{s_p\sqrt{\frac{2}{N}}} \sim t_{2N-2,\lambda}$$

where

$$\lambda = \Delta\sqrt{N/2}$$

# Two Sample Test, $\sigma$ 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, $\sigma$ 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, $\sigma$ Unknown

- Given $\beta$, 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, $\sigma$ 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, $\sigma$ 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

$$H_0 : \pi_1 = \pi_2 \quad \text{versus} \quad H_A : \pi_1 \neq \pi_2$$

- 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 $\delta_A$ using a two-sided test is

$$\Pr[Z > z_{1-\alpha/2} \mid \delta_A] + \Pr[Z < z_{\alpha/2} \mid \delta_A]$$
$$\approx \Pr[Z > z_{1-\alpha/2} \mid \delta_A]$$

- We need to know the distribution of $Z$ under $H_A$

$$E(p_2 - p_1) = \delta_A$$

$$\mathrm{Var}(p_2 - p_1) = \frac{\pi_2(1-\pi_2)}{N} + \frac{\pi_1(1-\pi_1)}{N}$$

# Two Sample Test: Binary Outcome

- So

$$
\begin{aligned}
1-\beta &= \Pr\left[\frac{p_2-p_1}{\sqrt{\frac{2\bar{p}\bar{q}}{N}}} > z_{1-\alpha/2} \mid \delta_A\right] \\
&= \Pr\left[p_2-p_1 > z_{1-\alpha/2}\sqrt{\frac{2\bar{p}\bar{q}}{N}} \mid \delta_A\right] \\
&= \Pr\left[\frac{(p_2-p_1)-\delta_A}{\sqrt{\mathrm{Var}(p_2-p_1)}} > \frac{z_{1-\alpha/2}\sqrt{\frac{2\bar{p}\bar{q}}{N}}-\delta_A}{\sqrt{\mathrm{Var}(p_2-p_1)}} \mid \delta_A\right]
\end{aligned}
$$

# Two Sample Test: Binary Outcome

- Implying

$$-z_{1-\beta} = \frac{z_{1-\alpha/2}\sqrt{2\bar{p}\bar{q}/N} - \delta_A}{\sqrt{\mathrm{Var}(p_2 - p_1)}}$$

- Using $\bar{p}\bar{q} \approx \bar{\pi}(1-\bar{\pi})$ where $\bar{\pi} = (\pi_1 + \pi_2)/2$ yields

$$z_{1-\beta}\sqrt{\mathrm{Var}(p_2 - p_1)} + z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})/N} = \delta_A$$

# 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 $\delta_A$ with power $1-\beta$ is

$$\frac{\left(z_{1-\beta}\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}+z_{1-\alpha/2}\sqrt{2\bar{\pi}(1-\bar{\pi})}\right)^2}{\delta_A^2}$$

# 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 $\pi_1$ against $\pi_2$ 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 | $\pi_{11}$ | $\pi_{12}$ |
| Unexposed | $\pi_{21}$ | $\pi_{22}$ |

# Case-Control: Binary Exposure

$$
\begin{aligned}
\text{OR} &= \frac{\text{odds}(\text{disease} \mid \text{exposed})}{\text{odds}(\text{disease} \mid \text{unexposed})} \\
&= \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} \\
&= \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} \\
&= \frac{\text{odds}(\text{exposed} \mid \text{disease})}{\text{odds}(\text{exposed} \mid \text{no disease})}
\end{aligned}
$$

# 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 $\pi_2$ we can determine $\pi_1$

$$\pi_1 = \frac{\pi_2 \, \text{OR}}{1 + \pi_2(\text{OR} - 1)}$$

# 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 $\text{OR} = 1.5$

## Case-Control, Binary Exposure: Example cont.

- First determine the corresponding value of $\pi_1$:

$$\pi_1 = \frac{0.2(1.5)}{1 + 0.2(0.5)} = 0.2727$$

- 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

$$Y_{1i} = \mu_i + \delta + \epsilon_{1i}; \quad i = 1, \ldots, N$$

- Controls ($k$ for each case)

$$Y_{2ij} = \mu_i + \epsilon_{2ij}; \quad j = 1, \ldots, k$$

- Assume $\epsilon_{1i}, \epsilon_{2ij}$ iid with

$$E(\epsilon_{1i}) = E(\epsilon_{2ij}) = 0$$

$$\mathrm{Var}(\epsilon_{1i}) = \mathrm{Var}(\epsilon_{2ij}) = \sigma^2$$

# 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

$$\mathrm{Var}(\bar{Y}_1) = \frac{\sigma^2}{N} \quad \text{and} \quad \mathrm{Var}(\bar{Y}_2) = \frac{\sigma^2}{kN}$$

- Therefore

$$\mathrm{Var}(\hat{\delta}_k) = \frac{\sigma^2}{N}\left(\frac{k+1}{k}\right)$$

- For $k = 1$,

$$\mathrm{Var}(\hat{\delta}_1) = \frac{2\sigma^2}{N}$$

# Case-Control: Continuous Exposure

- Relative efficiency

$$\text{eff}(\hat{\delta}_1, \hat{\delta}_k) = \frac{\text{Var}(\hat{\delta}_k)}{\text{Var}(\hat{\delta}_1)} = \frac{k+1}{2k} \to \frac{1}{2} \quad \text{as} \quad k \to \infty$$

| $k$ | $\text{eff}(\hat{\delta}_1, \hat{\delta}_k)$ |
|---|---|
| 1 | 1.00 |
| 2 | 0.75 |
| 3 | 0.67 |
| 4 | 0.63 |
| 5 | 0.60 |
| 10 | 0.55 |
| $\infty$ | 0.50 |

# Case-Control: Continuous Exposure

- Assuming $N$ large or $\epsilon_{1i}, \epsilon_{2ij} \sim N(0, \sigma^2)$
- Under $H_0 : \delta = 0$

$$Z = \frac{\hat{\delta}_k}{\sqrt{\mathrm{Var}(\hat{\delta}_k)}} \sim N(0, 1)$$

- Under $H_A : \delta = \delta_A > 0$,

$$1 - \beta = \Pr\left[\frac{\hat{\delta}_k - \delta_A}{\sqrt{\mathrm{Var}(\hat{\delta}_k)}} > z_{1-\alpha/2} - \frac{\delta_A}{\sqrt{\mathrm{Var}(\hat{\delta}_k)}}\right]$$

# Case-Control: Continuous Exposure

- Implying

$$-z_{1-\beta} = z_{1-\alpha/2} - \frac{\delta_A}{\sqrt{\mathrm{Var}(\hat{\delta}_k)}}$$

$$(z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{\delta_A^2}{\mathrm{Var}(\hat{\delta}_k)} = \delta_A^2 \frac{Nk}{\sigma^2(k+1)}$$

$$N = \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta_A^2} \frac{(k+1)}{2k}$$

# 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 $\alpha$ and $\beta$
- 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 $\alpha$
  - 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: $\sigma$ 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(mdiff=,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 + &mdiff; 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(mdiff=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 Frequency | Cumulative 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

$$\bar{Y}_i \sim N(\mu_i, \sigma^2/N_i) \quad \text{for} \quad i = 1, 2$$

- 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

$$\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{\sigma^2} \sim \chi^2_{N_1+N_2-2}$$

- 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 $N_1$, $N_2$, $\alpha$ and $\delta_A$, 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 $\Delta$
- 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 $\rho$ 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

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\left(Z_{\rho_1} - Z_{\rho_0}\right)^2} + 3$$

## Correlation Coefficient / Linear Regression

- For a simple linear regression model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

we have

$$\widehat{\beta}_1 = \frac{s_Y}{s_X} r$$

or

$$r = \frac{s_X}{s_Y} \widehat{\beta}_1$$

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

$$H_A : \beta_1 = 0.5 \quad \text{against} \quad H_0 : \beta_1 = 0$$

as to test

$$H_A : \rho = 0.1 \quad \text{against} \quad H_0 : \rho = 0$$

# 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

$$\begin{aligned} n &= \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(Z_{\rho_1} - Z_{\rho_0})^2} + 3 \\ &= \frac{(1.96 + 1.28)^2}{(0.1003 - 0)^2} + 3 \\ &= 1046 \end{aligned}$$

# 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, $X_2, X_3, X_4, \ldots$
- 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

$$n' = \text{VIF} \cdot n = \frac{1}{1 - R^2} \cdot n = \frac{1}{1 - r^2} \cdot n$$

$$= \frac{1}{1 - 0.3^2} \cdot 1046 = 1149$$

# 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 $k^{\text{th}}$ person in the $j^{\text{th}}$ cluster at the $i^{\text{th}}$ treatment level,

$$i = 1, 2, \ldots, I; \quad j = 1, 2, \ldots, J; \quad k = 1, 2, \ldots, K$$

- 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

$$
\begin{aligned}
\mathrm{Var}(\bar{Y}_{ij}) &= E(\bar{Y}_{ij}^2) - \mu_i^2 \\
&= K^{-2} E\left(\sum_{k=1}^{K} Y_{ijk}\right)^2 - \mu_i^2 \\
&= K^{-2} E\left(\sum_{k=1}^{K} Y_{ijk}^2 + \sum\sum_{k \neq k'} Y_{ijk} Y_{ijk'}\right) - \mu_i^2 \\
&= K^{-2}\Big(K\sigma^2 + K(K-1)\rho\sigma^2\Big) \\
&= \frac{\sigma^2}{K}\Big(1 + (K-1)\rho\Big)
\end{aligned}
$$

# Variance Inflation Factor (VIF)

- $\big(1 + (K-1)\rho\big)$ is the *variance inflation factor* (VIF)
- It measures the increase in the variance of the mean due to the within-subject correlation of measurements ($\rho$)
- 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

$$\mathrm{Var}(\bar{Y}_i) = \frac{\sigma^2}{JK}\mathrm{VIF}$$

# 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:

| | $\rho$ | | | | |
|---|---|---|---|---|---|
| $K$ | 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 $k = 1, \ldots, K_j$
- 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 ($\pi$) 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 $\pi$ 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 $\pi$ 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

$$\hat{p}_{\mathrm{u}} = 330/1320 = 0.25; \quad \hat{p}_{\mathrm{r}} = 414/1338 = 0.31$$

- 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

| | Urban | | | Rural | | |
|---|---|---|---|---|---|---|
| Age | $N_{1k}$ | $n_{1k}$ | $\hat{p}_{1k}$ | $N_{2k}$ | $n_{2k}$ | $\hat{p}_{2k}$ |
| 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)

$$\hat{p}_{1_{\text{adj}}} = (0.194 \times 0.062 + \cdots + 0.256 \times 0.285)/1 = 0.224$$

$$\hat{p}_{2_{\text{adj}}} = 0.354$$

- 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:

$$\hat{p}_{1_{\text{adj}}} = \frac{6 \times 0.062 + \cdots + 5 \times 0.285}{6 + 6 + 6 + 5} = 0.209$$

$$\hat{p}_{2_{\text{adj}}} = 0.330$$

$$\hat{p}_{1_{\text{adj}}} - \hat{p}_{2_{\text{adj}}} = -0.121$$

$$\frac{\hat{p}_{2_{\text{adj}}}}{\hat{p}_{1_{\text{adj}}}} = 1.58$$

# 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

$$\widehat{\mathrm{Var}}(\hat{p}_{1_{\mathrm{adj}}} - \hat{p}_{2_{\mathrm{adj}}}) = \frac{\sum_{k=1}^{K} w_k^2 \left(\widehat{\mathrm{Var}}(\hat{p}_{1k}) + \widehat{\mathrm{Var}}(\hat{p}_{2k})\right)}{\left(\sum_{k=1}^{K} w_k\right)^2}$$

- Large sample tests and CIs are obtained from the CLT

# Direct Standardization

- Revisiting the smoking example (using combined weights)

$$\widehat{\mathrm{Var}}(\hat{p}_{1_{\mathrm{adj}}} - \hat{p}_{2_{\mathrm{adj}}}) = 0.000318$$

- Testing $H_0 : \pi_{1_{\mathrm{adj}}} = \pi_{2_{\mathrm{adj}}}$ versus $H_A : \pi_{1_{\mathrm{adj}}} \neq \pi_{2_{\mathrm{adj}}}$,

$$Z = \frac{\hat{p}_{1_{\mathrm{adj}}} - \hat{p}_{2_{\mathrm{adj}}}}{\sqrt{\widehat{\mathrm{Var}}(\hat{p}_{1_{\mathrm{adj}}} - \hat{p}_{2_{\mathrm{adj}}})}} = \frac{-0.13}{\sqrt{0.000318}} = -7.28$$

- 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 $k = 1, \ldots, K$
  - Study population: $n_k/N_k$ for $k = 1, \ldots, K$
- 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

$$\widehat{\mathrm{Var}}(s) = \frac{\widehat{\mathrm{Var}}(O) + s^2\widehat{\mathrm{Var}}(E)}{E^2}$$

where $\widehat{\mathrm{Var}}(O) = \sum_k n_k$

and $\widehat{\mathrm{Var}}(E) = \sum_k \left(\frac{N_k}{M_k}\right)^2 m_k$

- To test $H_0 : \pi_{\text{study}}/\pi_{\text{ref}} = 1$ vs. $H_0 : \pi_{\text{study}}/\pi_{\text{ref}} \neq 1$,

$$Z = \frac{s-1}{\sqrt{\widehat{\mathrm{Var}}(s)}} \sim N(0,1)$$

# 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 $\widehat{\mathrm{Var}}(O) = O = 414$ and

$$\begin{aligned}\widehat{\mathrm{Var}}(E) &= 8\left(\frac{387}{129}\right)^2 + 53\left(\frac{441}{243}\right)^2 + 135\left(\frac{300}{478}\right)^2 + 134\left(\frac{210}{470}\right)^2 \\ &= 326.49\end{aligned}$$

- Therefore

$$\begin{aligned}\widehat{\mathrm{Var}}(s) &= \frac{\widehat{\mathrm{Var}}(O) + s^2\widehat{\mathrm{Var}}(E)}{E^2} = \frac{414 + 1.56^2(326.49)}{264.79^2} \\ &= 0.0173\end{aligned}$$

implying

$$Z = \frac{s-1}{\sqrt{\widehat{\mathrm{Var}}(s)}} = 4.29$$

# 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:

$$S(0) = 1; \quad S(\infty) = 0$$

$$\text{If } t_1 \leq t_2, \text{ then } S(t_1) \geq S(t_2)$$

# 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 $i^{\text{th}}$ 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 $(T_i, \delta_i)$ for $i = 1, 2, \ldots, N$

# Example

- Remission time in weeks for leukemia patients ($N = 21$)

| $(T_i, \delta_i)$ | $(T_i, \delta_i)$ | $(T_i, \delta_i)$ |
|---|---|---|
| (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 $t_{(1)}, t_{(2)}, \ldots, t_{(J)}$ be the distinct ordered failure times (censoring times are ignored)

| Failure time $t_{(j)}$ | Risk set $R(t_{(j)})$ | No. of failures $m_j$ | No. censored in $[t_{(j)}, t_{(j+1)})$ $q_j$ |
|---|---|---|---|
| $t_{(0)} = 0$ | $R(t_{(0)}) = N$ | $m_0 = 0$ | $q_0$ |
| $t_{(1)}$ | $R(t_{(1)})$ | $m_1$ | $q_1$ |
| $t_{(2)}$ | $R(t_{(2)})$ | $m_2$ | $q_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $t_{(J)}$ | $R(t_{(J)})$ | $m_J$ | $q_J$ |

- $R(t_{(j)}) = R(t_{(j-1)}) - m_{j-1} - q_{j-1}$

# 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 $t \in [t_{(j)}, t_{(j+1)})$

$$\begin{aligned}\hat{S}(t) &= \hat{S}(t_{(j-1)}) \cdot \widehat{\Pr}[T > t_{(j)} | T \geq t_{(j)}] \\ &= \hat{S}(t_{(j-1)}) \left( \frac{R(t_{(j)}) - m_j}{R(t_{(j)})} \right)\end{aligned}$$

- 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)
> 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

| Total | Failed | Censored | Percent Censored |
|---|---|---|---|
| 21 | 9 | 12 | 57.14 |

# Greenwood SE/CI of KM

- Let $n_j = R(t_{(j)})$
- Write the Kaplan-Meier estimator as

$$\hat{S}(t) = \prod_{j=1}^{i} \hat{p}_j \quad \text{for} \quad t \in [t_{(i)}, t_{(i+1)}),$$

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

$$\mathrm{Var}\big(\log \hat{S}(t)\big) = \sum_{j=1}^{i} \mathrm{Var}\big(\log \hat{p}_j\big)$$

- Binomial argument

$$\widehat{\mathrm{Var}}(\hat{p}_j) = \hat{p}_j(1 - \hat{p}_j)/n_j$$

# Greenwood SE/CI for KM

- Taylor series approximation

$$\widehat{\mathrm{Var}}\big(g(X)\big) \approx \big(g'(\mu)\big)^2 \widehat{\mathrm{Var}}(X)$$

implies

$$\widehat{\mathrm{Var}}(\log \hat{p}_j) \approx \left(\frac{1}{\hat{p}_j}\right)^2 \left(\frac{\hat{p}_j(1-\hat{p}_j)}{n_j}\right) = \frac{1-\hat{p}_j}{n_j\,\hat{p}_j}$$

$$= \frac{m_j}{n_j(n_j - m_j)}$$

- Thus

$$\widehat{\mathrm{Var}}\big(\log \hat{S}(t)\big) \approx \sum_{j=1}^{i} \frac{m_j}{n_j(n_j - m_j)}$$

# Greenwood SE/CI for KM

- Additional application of Taylor series approximation

$$\widehat{\mathrm{Var}}\big(\log \hat{S}(t)\big) \approx \big(\hat{S}(t)\big)^{-2}\widehat{\mathrm{Var}}\big(\hat{S}(t)\big)$$

implying

$$\widehat{\mathrm{Var}}\big(\hat{S}(t)\big) \approx \big(\hat{S}(t)\big)^{2}\sum_{j=1}^{i}\frac{m_j}{n_j(n_j-m_j)}$$

- Thus

$$\widehat{\mathrm{SE}}(\hat{S}(t)) \approx \hat{S}(t)\sqrt{\textstyle\sum_{j=1}^{i}\frac{m_j}{n_j(n_j-m_j)}}$$

for $t_{(i)} \leq t < t_{(i+1)}$

# Greenwood SE/CI for KM

- For the leukemia example,

$$\widehat{\mathrm{SE}}(\hat{S}(6)) = 0.8571\sqrt{\frac{3}{21 \cdot (21-3)}} = 0.0764$$

$$\widehat{\mathrm{SE}}(\hat{S}(7)) = 0.8067\sqrt{\frac{3}{21 \cdot 18} + \frac{1}{17 \cdot 16}} = 0.0869$$

- An approximate $100(1-\alpha)\%$ CI is

$$\hat{S}(t) \pm z_{1-\alpha/2}\,\widehat{\mathrm{SE}}(\hat{S}(t))$$

# 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 $j = 1, 2, \ldots, n_i$

- We want to test

$$H_0 : S_1(t) = S_2(t) \quad \text{for all } t$$

where

$$S_j(t) = \Pr[T_j^* > t] \quad \text{for} \quad j = 1, 2$$

# Log-Rank Test

- Let $t_{(1)}, t_{(2)}, \ldots, t_{(K)}$ 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

$$V_{1k} = \frac{R_1(t_{(k)}) R_2(t_{(k)}) m_k \big(R(t_{(k)}) - m_k\big)}{R(t_{(k)})^2 \left(R(t_{(k)}) - 1\right)}$$

# 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 |
| | 9 | | | | | | 14.57 | 5.4 |

# 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 $2 \times 2$ 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 $t$

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)}$$

- 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 $x_2 = 0$,

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{e^{\beta \cdot 1}}{e^{\beta \cdot 0}} = e^{\beta}$$

- $e^{\beta}$ 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,

$$\mathrm{Var}(\bar{y}) = \frac{1}{n^2}\left(\sum_{i=1}^{N} y_i^2 \mathrm{Var}(Z_i) + \sum_{i \neq j} y_i y_j \mathrm{Cov}(Z_i, Z_j)\right)$$

  we need the variance and covariance terms

- The variance is easy because the $Z_i$ are Bernoulli

$$\mathrm{Var}(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$

- For SRS, the $Z_i$ are not independent

$$\begin{aligned}\mathrm{Cov}(Z_i, Z_j) &= E(Z_i Z_j) - E(Z_i)E(Z_j) = \frac{n}{N}\frac{(n-1)}{(N-1)} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{n}{N}\left(1 - \frac{n}{N}\right)\frac{1}{N-1}\end{aligned}$$

# SRS: Estimating Population Mean

- Thus

$$\mathrm{Var}(\bar{y}) = \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i\neq j} y_i y_j\right)$$

- 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

$$\mathrm{Var}(\bar{y}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{\sum(y_i-\mu)^2}{N-1} = \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}$$

# 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

$$\mathrm{Var}(\bar{y}) \approx \frac{\sigma^2}{n}$$

- On the other hand, $\mathrm{Var}(\bar{y}) \to 0$ as $n \to N$

# 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 $\mathrm{Var}(\bar{y})$ is given by

$$\widehat{\mathrm{Var}}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$$

# 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

$$\mathrm{Var}(\hat{\tau}) = N^2\mathrm{Var}(\bar{y}) = N(N-n)\frac{\sigma^2}{n}$$

- Unbiased estimator of variance

$$\widehat{\mathrm{Var}}(\hat{\tau}) = N^2\widehat{\mathrm{Var}}(\bar{y}) = N(N-n)\frac{s^2}{n}$$

# 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 $\mu_N$ be the population mean and $\bar{y}_N$ the sample mean for an SRS from that population.
- According to the finite-population CLT

$$\frac{\bar{y}_N - \mu_N}{\sqrt{\mathrm{Var}(\bar{y}_N)}} \to Z \sim N(0,1)$$

as both $n \to \infty$ and $N - n \to \infty$

# SRS: Finite-population CIs

- This leads to approximate $100(1-\alpha)\%$ CIs for the population mean $\mu$

$$\bar{y} \pm t_{n-1,1-\alpha/2}\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

- Likewise, for the population total $\tau$

$$\hat{\tau} \pm t_{n-1,1-\alpha/2}\sqrt{N(N-n)\frac{s^2}{n}}$$

# 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

$$\widehat{\mathrm{Var}}(\bar{y}) = \left(1 - \frac{15}{286}\right)\frac{919.067}{15} = 58.058$$

yielding 95% CI for $\mu$

$$25.933 \pm 2.145\sqrt{58.058} = (9.59, 42.28)$$

- 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;
```

```
                          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    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


> svytotal(caribou$y, design); confint(svytotal(caribou$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

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \sim N(0, 1)$$

choose $n$ such that

$$z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})} = d$$

# 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 $N >> n$, 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 $\mu$ 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

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - p)^2}{N-1} = \frac{\sum_{i=1}^{N} y_i^2 - Np^2}{N-1}$$

$$= \frac{Np - Np^2}{N-1}$$

$$= \frac{Np(1-p)}{N-1}$$

# SRS: Estimating a Proportion

- Proportion in the sample with the attribute

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{N} y_i Z_i = \overline{y}$$

- 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:

$$E(\hat{p}) = p$$

$$\mathrm{Var}(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$$

$$\widehat{\mathrm{Var}}(\hat{p}) = \left(\frac{N-n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$$

$$E(\widehat{\mathrm{Var}}(\hat{p})) = \mathrm{Var}(\hat{p})$$

# SRS: CI for a Proportion

- Approximate $100(1-\alpha)\%$ CI

$$\hat{p} \pm t_{1-\alpha/2,n-1}\sqrt{\widehat{\mathrm{Var}}(\hat{p})}$$

- 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 $\nu$ 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 $\nu_L$ be the smallest integer such that

$$\Pr[X \geq x | \nu_L] > \alpha/2$$

and let $\nu_U$ 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

$$n = \frac{Np(1-p)}{(N-1)d^2/z^2_{1-\alpha/2} + p(1-p)}$$

- If $N >> n$

$$n \approx \frac{z^2_{1-\alpha/2}\, p(1-p)}{d^2}$$

- 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 $h = 1, \ldots, H$.
- 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 $h$ $(i = 1, \ldots, N_h;\ h = 1, \ldots, H)$
- 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 $\mu_h$ for stratum $h$

- $E(\bar{y}_h) = \mu_h$ implies $E(\bar{y}) = \mu$

- Estimator of the variance of $\bar{y}$

$$\widehat{\mathrm{Var}}(\bar{y}) = \sum_h W_h^2 \widehat{\mathrm{Var}}(\bar{y}_h)$$

- $E(\widehat{\mathrm{Var}}(\bar{y}_h)) = \mathrm{Var}(\bar{y}_h)$ implies $E(\widehat{\mathrm{Var}}(\bar{y})) = \mathrm{Var}(\bar{y})$

# Population Mean Estimator

- If stratified random sampling, then

$$\bar{y}_h = \frac{\sum_i y_{hi} Z_{hi}}{n_h}$$

and

$$\widehat{\mathrm{Var}}(\bar{y}) = \sum_h W_h^2 \left( \frac{1 - f_h}{n_h} \right) s_h^2$$

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

$$\bar{y} \pm t_{1-\alpha/2,df}\sqrt{\widehat{\mathrm{Var}}(\bar{y})}$$

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

$$\widehat{\mathrm{Var}}(\hat{\tau}) = N^2 \widehat{\mathrm{Var}}(\bar{y}) = \sum_h N_h^2 \left( \frac{1 - f_h}{n_h} \right) s_h^2$$

with the second equality assuming stratified random sampling

# Population Total Estimator

- $E(\widehat{\mathrm{Var}}(\bar{y}_h)) = \mathrm{Var}(\bar{y}_h)$ implies $E(\widehat{\mathrm{Var}}(\hat{\tau})) = \mathrm{Var}(\hat{\tau})$

- CIs

$$\hat{\tau} \pm t_{1-\alpha/2,df}\sqrt{\widehat{\mathrm{Var}}(\hat{\tau})}$$

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

$$\widehat{\mathrm{Var}}(\hat{p}) = \sum_h W_h^2 \left( \frac{1 - f_h}{n_h - 1} \right) \hat{p}_h (1 - \hat{p}_h)$$

## 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 $n = 4$,

$$\mathrm{Var}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n} = \left(1 - \frac{4}{6}\right)\frac{10.7}{4} = 0.89$$

- For stratified random sampling with $n_1 = n_2 = 2$,

$$\mathrm{Var}(\bar{y}) = \left(\frac{3}{6}\right)^2\left(\frac{1 - 2/3}{2}\right)1 + \left(\frac{3}{6}\right)^2\left(\frac{1 - 2/3}{2}\right)7 = 0.33$$

# 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

$$\mathrm{Var}_{\mathrm{SRS}}(\bar{y}) = (1-f)\,\sigma^2/n$$

$$\approx (1-f)\sum_h W_h\sigma_h^2/n + (1-f)\sum_h W_h(\mu_h-\mu)^2/n$$

- Under proportionate stratified random sampling

$$\mathrm{Var}_{\mathrm{pro}}(\bar{y}) = (1-f)\sum_h W_h^2\sigma_h^2/n_h = (1-f)\sum_h W_h\sigma_h^2/n$$

- Therefore

$$\mathrm{Var}_{\mathrm{SRS}}(\bar{y}) \approx \mathrm{Var}_{\mathrm{pro}}(\bar{y}) + (1-f)\sum_h W_h(\mu_h-\mu)^2/n$$

# 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:
  - $\mu$, the average number of hours of certification training in the year prior to the last certification
  - $\mu_1$, the average number of hours of certification training in City A in the year prior to the last certification
  - $\tau$, 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 | 1 | 8 | 1 |
| 2 | 28 | 1 | 12 | 31 | 1 | 2 | 4 | 0 | 2 | 15 | 0 |
| 3 | 26 | 1 | 13 | 39 | 1 | 3 | 49 | 0 | 3 | 21 | 1 |
| 4 | 41 | 1 | 14 | 38 | 0 | 4 | 10 | 1 | 4 | 7 | 0 |
| 5 | 43 | 1 | 15 | 40 | 0 | 5 | 15 | 0 | 5 | 14 | 1 |
| 6 | 29 | 0 | 16 | 45 | 1 | 6 | 41 | 0 | 6 | 30 | 1 |
| 7 | 32 | 1 | 17 | 28 | 1 | 7 | 25 | 0 | 7 | 20 | 0 |
| 8 | 37 | 1 | 18 | 27 | 1 | 8 | 30 | 0 | 8 | 11 | 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 $\mu$, the average number of hours of certification training in the year prior to the last certification.
- Estimate

$$\bar{y} = \sum_h W_h \bar{y}_h$$

$$= 0.5 \times 33.900 + 0.2 \times 25.125 + 0.3 \times 19.000$$

$$= 27.675 \text{ hours}$$

# Example

- Estimated variance

$$\widehat{\mathrm{Var}}(\bar{y}) = \sum_h W_h^2 \left(\frac{1-f_h}{n_h}\right) s_h^2$$

$$= 0.5^2 \left(\frac{1-0.129}{20}\right) 35.358$$

$$+ 0.2^2 \left(\frac{1-0.129}{8}\right) 232.411$$

$$+ 0.3^2 \left(\frac{1-0.129}{12}\right) 87.636$$

$$= 1.97$$

# 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

                                        Std Error
  Variable         N            Mean      of Mean       95% CL for Mean
  ---------------------------------------------------------------------
  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: $\mu_p = \dfrac{\tau}{N}$
- Population mean per secondary unit: $\mu = \dfrac{\tau}{M}$
- 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 $\tau$ 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

$$\mathrm{Var}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n}$$

where $\sigma_u^2$ 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

$$\widehat{\mathrm{Var}}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n}$$

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 $y_1, \ldots, y_N$

- An unbiased estimator of $\mu_p$ is given by $\bar{y} = \hat{\tau}/N$
- An unbiased estimator of $\mu$ 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 $k^{\text{th}}$ 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 $i^{\text{th}}$ selected primary unit, for $i = 1, \ldots, n$

- $\mu_i = y_i/M_i$ is the mean per secondary unit in the $i^{\text{th}}$ primary unit
- $Z_i$ is as before
- $Z_{ij} = 1$ if the $j^{\text{th}}$ secondary unit of the $i^{\text{th}}$ primary unit is in the sample, 0 otherwise

# Two-Stage Cluster Sampling

- If the $i^{\text{th}}$ 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

$$E(\hat{\tau}) = E\big(E(\hat{\tau}|Z_1, \ldots, Z_n)\big)$$

# Two-Stage Cluster Sampling

- First evaluate the inner expectation

$$E(\hat{\tau}|Z_1,\ldots,Z_n) = E\left(\frac{N}{n}\sum_{i=1}^{N}\hat{y}_i Z_i \middle| Z_1,\ldots,Z_N\right)$$

$$= \frac{N}{n}\sum_{i=1}^{N} E(\hat{y}_i|Z_i=1)Z_i$$

$$= \frac{N}{n}\sum_{i=1}^{N} y_i Z_i$$

- 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

$$\mathrm{Var}(\hat{\tau}) = N(N-n)\frac{\sigma_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$$

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 $i = 1, \ldots, N$

# Two-Stage Cluster Sampling

- Note that the first term in $\mathrm{Var}(\hat{\tau})$ 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 $\mathrm{Var}(\hat{\tau})$ 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 $\mathrm{Var}(\hat{\tau})$, we will use the fact

$$\mathrm{Var}(\hat{\tau}) = \mathrm{Var}\big(E(\hat{\tau}|Z_1,\ldots,Z_N)\big) + E\big(\mathrm{Var}(\hat{\tau}|Z_1,\ldots,Z_N)\big)$$

- For the first term, we have

$$\mathrm{Var}\big(E(\hat{\tau}|Z_1,\ldots,Z_N)\big) = \mathrm{Var}\left(\frac{N}{n}\sum_{i=1}^{N} y_i Z_i\right) = N(N-n)\frac{\sigma_u^2}{n}$$

where the second equality follows from results for SRS

- To evaluate the second term, first note that

$$\mathrm{Var}(\hat{y}_i|Z_i = 1) = M_i(M_i - m_i)\frac{\sigma_i^2}{m_i}$$

# Two-Stage Cluster Sampling

- Therefore

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}|Z_1,\ldots,Z_n) &= \mathrm{Var}\left(\frac{N}{n}\sum_{i=1}^{N}\hat{y}_i Z_i \middle| Z_1,\ldots,Z_N\right) \\
&= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N} \mathrm{Var}(\hat{y}_i|Z_i=1)Z_i \\
&= \left(\frac{N}{n}\right)^2 \sum_{i=1}^{N} M_i(M_i-m_i)\frac{\sigma_i^2}{m_i}Z_i
\end{aligned}$$

- Thus

$$E\big(\mathrm{Var}(\hat{\tau}|Z_1,\ldots,Z_n)\big) = \frac{N}{n}\sum_{i=1}^{N} M_i(M_i-m_i)\frac{\sigma_i^2}{m_i}$$

# Two-Stage Cluster Sampling

- An unbiased estimator of the variance of $\hat{\tau}$ is

$$\widehat{\mathrm{Var}}(\hat{\tau}) = N(N-n)\frac{s_u^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i(M_i - m_i)\frac{s_i^2}{m_i} Z_i$$

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 $i = 1, \ldots, N$

- The proof is left as an exercise

# Two-Stage Cluster Sampling

- Estimators for population means follow immediately:

$$\hat{\mu}_p = \hat{\tau}/N \quad \text{is unbiased for} \quad \mu_p$$

$$\hat{\mu} = \hat{\tau}/M \quad \text{is unbiased for} \quad \mu$$

- Variance expressions follow from $\mathrm{Var}(\hat{\tau})$ 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

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\hat{\tau}) &= 100(100-3)\frac{17100}{3} \\
&\quad + \frac{100}{3}\left(24(24-2)\frac{8}{2} + 20(20-2)\frac{0}{2} + 15(15-2)\frac{2}{2}\right) \\
&= 55{,}366{,}900
\end{aligned}
$$

# 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, disproportionate, balanced*
- Cluster
  - *One stage* – SRS of clusters; sample all within cluster
  - *Systematic (list)*
  - *Multi-stage*, for example, blocks then dwellings

# ERIC Notebook

Second Edition

## Assessment of Diagnostic and Screening Tests


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


Diagnostic and screening tests are used to detect the presence or severity of disease in individuals. Clinicians rely on these tests to make decisions in treating patients. Thus, it is important to assess the performance of diagnostic and screening tests before they are adopted in a clinical setting. The performance of any new diagnostic or screening test is assessed by comparing actual test results to the patient's true disease status (as assessed by a *gold standard*). The four measures used to evaluate a new test are the *sensitivity*, *specificity,* and *positive and negative predictive values.*

### The gold standard

Gold standard is a term for the most definitive diagnostic procedure, e.g. microscopic examination of a tissue specimen, or the best available laboratory test, e.g. serum antibodies to HIV. Sometimes it can refer to a comprehensive clinical evaluation, e.g. clinical assessment of arthritis. Gold standard procedures can often be costly, invasive, and/or uncomfortable. New tests that are less invasive and less expensive are compared against gold standards to assess the new test's accuracy. A test which detects a marker in the blood for prostate cancer may not be as sensitive as taking a biopsy from the prostate itself, but the discomfort of biopsy may make the blood assay a better alternative.

### Calculating the test results

A table like the one below is used to group individuals into one of four disease-test categories.

| | Truth (Gold Standard) | |
|---|---|---|
| Results of test | Disease Present | Disease Absent |
| Positive Test | a = True Positive | b = False Positive |
| Negative Test | c = False Negative | d = True Negative |

Prevalence of the disease =

$$\frac{\text{True positives}}{\text{Total population}} = \frac{(a+c)}{a+b+c+d)}$$

### Sensitivity vs. specificity

Sensitivity and specificity are measures that assess the validity of diagnostic and screening tests. These measures reflect how well the test is

detecting the disease and classifying individuals into disease and non-disease groups.

**Sensitivity (Se or Sn)** describes how well the test detects disease in all who truly have disease, or the percent of diseased individuals who have positive test results

**Specificity (Sp),** describes how well the test is detecting non-diseased individuals as truly not having the disease, or the percent of non-diseased individuals who have

$$\text{Se} = \frac{True-positives}{True-positivies+False-negatives} \times 100$$

$$\text{Se} = \frac{a}{(a+c)} \times 100$$

$$\text{Sp} = \frac{True-negatives}{True-negativies+False-positives} \times 100$$

$$\text{Sp} = \frac{d}{(b+d)} \times 100$$

A highly sensitive test means that a large percent of people who have disease are classified correctly as having the disease. A highly specific test means that a large percent of individuals without disease are classified correctly as not having disease. An ideal test would be both highly sensitive and highly specific, where disease would be detected in 100% of those who truly have disease (100% sensitivity), and disease would be ruled out in 100% of those who are truly disease-free (100% specificity).

**Example**

If a test is 95% sensitive and 98% specific, then 5% of the diseased individuals will have negative test results (the test is incorrectly classifying 5% of the diseased individuals), and 2% of the disease-free individuals tested will have positive test results (the test is incorrectly classifying 2% of the disease-free individuals).

### False-positives and false-negatives

A *false positive* is an individual who is incorrectly diagnosed as a case when, in fact, they do not have the disease. A *false negative* is an individual who is incorrectly diagnosed as a non-case, when in fact the person does have the disease.

100% - % sensitivity = % false negatives

100% - % specificity = % false positives

### Positive and negative predictive values

The *positive predictive value* (PPV) is the percent of positive tests that are truly positive. The *negative predictive value* (NPV) is the percent of negative tests that are truly negative.

Like sensitivity and specificity, PPV and NPV also show how well the test is classifying individuals into disease and non-disease groups, but the denominator for PPV is the total number of persons who test positive (a+b), while that for NPV is the total number who test negative (c+d). A test with a high PPV value means that there is only a small percent of false-positives within all the individuals with positive test results. A test with a high NPV value means that there is only a small percent of false-negatives within all the individuals with negative test results.

$$\text{PPV} = \frac{True\ Positives}{Positive\ Tests} \times 100$$

$$\text{PPV} = \frac{a}{(a+b)} \times 100$$

$$\text{NPV} = \frac{True\ Negatives}{Negative\ Tests}\ x100$$

$$\text{NPV} = \frac{d}{(c+d)} \times 100$$

**Example**

A certain test, e.g. a stress ECG, which has a PPV of 90% and a NPV of 95% is used to screen 5,000 people for coronary heart disease. Forty percent of the individuals (2,000 people) have positive test results and 60% (3,000 people) have negative test results. But the gold standard for CHD found that 1,800 of those who tested positive (90% of 2,000) truly have CHD, and 2,850 of those who tested negative (95% of 3,000) are truly non-cases.

### Pros and cons of specificity and sensitivity

Ideally, an investigator would prefer a diagnostic test that is both 100% sensitive and 100% specific. However, this scenario rarely occurs. It is important in clinical decision-making to know the sensitivity and specificity of the test you are conducting and to weigh the pros and cons of using tests with different levels of sensitivity and specificity. For instance, if a disease is not life threatening if left untreated, the costs of treatment are high, and invasive surgery is required, then a very specific diagnostic test is preferred over a more sensitive test. If the disease under study is life threatening if left untreated, and the survival rate is improved with immediate treatment, then the sensitivity of a diagnostic test is of greater importance than its specificity.

### Example

Schizophrenia has a low prevalence in the U.S. at around 1%. A new diagnostic test that is 99% sensitive and 99% specific is used to screen 10,000 patients for schizophrenia. Of those 10,000, we would expect 100 to truly be suffering from schizophrenia, or 1% of our population. Of those 100, 99 (99% of 100) would have positive test results. Of the 9,900 who are truly without disease, 9801 (99% of 9,900) would be classified as disease-free. However, there would be 99 (9,900-9801) false positives. This test would give 198 (99+99) positive test results. Therefore, even with a test that is 99% sensitive and 99% specific, the PPV would only be 50% (99/198).

### The prevalence affects the predictive values

The prevalence of a disease affects the PPV and NPV values. If a disease has a low prevalence and the test being used to assess disease in individuals is not 100% sensitive or 100% specific, as will most likely be the case, then false-positives may overwhelm the positive test results.

### Terminology

*False positive:* An individual who is incorrectly diagnosed as having disease

*False negative:* An individual who is incorrectly diagnosed as not having disease

*Gold standard:* The most definitive diagnostic procedure to detect disease

*Negative predictive value:* The percent of negative tests that are truly negative

*Positive predictive value:* The percent of positive tests that are truly positive

*Sensitivity:* The percent of diseased individuals who have positive test results

*Specificity:* The percent of non-diseased individuals who have negative test results

### Practice Questions.

**A test is used to screen people for hepatitis B. The sensitivity of the test is 95% and the specificity of the test is 90%. Assume that the total number of persons being tested for hepatitis B is 50,000. Assume that the true prevalence of hepatitis B in the population is 100 per 50,000.**

| | Disease present | Disease absent |
|---|---|---|
| Positive test | a= true positive | b= false positive |
| Negative test | c= false negative | d= true negative |

1) Calculate the number of true positives

2) Calculate the number of false positives

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

## Answers to Practice Questions

| | Disease present | Disease absent |
|---|---|---|
| Positive test | a= true positive | b= false positive |
| Negative test | c= false negative | d= true negative |

a+ b+ c+ d= 50,000
a+c= 100 (100 people have Hepatitis B)
b+d = Total – (a+c) = 50,000-100=49900

1) Calculate the number of true positives
Sensitivity = 95% = a / (a+c) a= (0.95)(a+c)= 0.95*100= 95 true positives

A sensitivity of 95% means that the test detects as positive 95% of the diseased individuals.

2) Calculate the number of false positives
Specificity = 90%= d / (d+b)
d=(0.90)(d+b) = 0.90*49900)=44910
b= (b+d)-d= (49900-44910)=4990 false positives

A specificity of 90% means that the test detects as negative 90% of the non-diseased individuals.

# ERIC Notebook

Second Edition

## Calculating Person-Time


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

### What is person-time?

Person-time is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a study. In certain studies people are followed for different lengths of time, as some will remain free of a health outcome or disease longer than others. A subject is eligible to contribute person-time to the study only so long as that person does not yet have the health outcome under study and, therefore, is still at risk of developing the health outcome of interest. By knowing the number of new cases of the health outcome and the person-time-at-risk contributed to the study, an investigator can calculate the rate of the health outcome or disease, or how quickly people are acquiring the health outcome or disease.

### Calculating rates

The rate is the number of new (incident) cases during study follow-up divided by the person-time-at-risk throughout the observation period.

$$\text{Rate} = \frac{\#\ of\ new\ cases}{total\ person\text{-}time\ at\ risk}$$

The denominator for a rate (person-time) is a more exact expression of the population at risk during the period of time when the change from non-disease to disease is being measured. The denominator for the rate changes as persons originally at risk develop the health outcome during the observation period and are removed from the denominator.

### Calculating person-time for rates

Now suppose an investigator is conducting a study of the rate of second myocardial infarction (MI). He follows 5 subjects from baseline (first MI) for up to 10 weeks. The results are graphically displayed as follows:

The graph shows how many days each subject remained in the study as a non-case (no second MI) from baseline. From this graph the investigator can calculate person-

time. Person-time is the sum of total time contributed by all subjects. The unit for person-time in this study is person-days (p-d).

Time contributed by each subject:

Subject A: 53 days

Subject B: 70 days

Subject C: 24 days

Subject D: 70 days

Subject E: 19 days

Total person-days in the study: 53+70+24+70+19=236 person-days

236 person-days (p-d) now becomes the denominator in the rate measure. The total number of subjects becoming cases (subjects A, C, and E) is the numerator in the rate measure. Therefore the rate of secondary MI is 3/(236 p-d), which is 0.0127 cases per person-day. By multiplying the numerator and denominator by 1000, the rate becomes 12.7 cases per 1000 person-days. The denominator, person-days, can be converted into other time units (such as hours or years) appropriate to the disease or health outcome being studied.

Secondary MI may be expressed in cases per person-year (p-y) by: (0.0127 cases/p-d) x (365 p-d/1 p-y) = 4.6 cases/p-y

**Estimating when a person becomes a case**

Now suppose an investigator is studying the rate of prostate cancer in men with a family history of prostate cancer. Subjects are examined once a year for up to five years. In order to calculate person-time when an investigator is only examining patients at specified intervals (once a year) the investigator must determine when a newly diagnosed case acquired the disease within the last year. In order to determine the amount of person-time adequately, an investigator may decide that the onset of prostate cancer occurred at the midpoint of the time interval between being disease free and becoming a case. This is because the investigator does not know precisely when subject A developed prostate cancer (just that it was sometime between exams two and three).

The following graph displays the amount of time until onset of prostate cancer for each subject.

Time contributed by each subject:

Subject A: 2.5 years

Subject B: 5 years

Subject C: 1.5 years

Subject D: 5 years

Subject E: 0.5 years

Total person-years in the study:

(2.5+5+1.5+5+0.5)=14.5 person-years

14.5 p-y is the denominator in the rate of prostate cancer. The rate is 3/(14.5 p-y), or 0.207 cases per p-y. By multiplying both the numerator and denominator by 1000 the rate becomes 207 cases per 1000 p-y.

**Terminology**

*Rate:* the number of new cases of disease during a period of time divided by the person-time-at-risk

*Person-time:* estimate of the actual time-at-risk in years, months, or days that all persons contributed to a study

## Practice Questions

*Answers are located at end of this notebook.*

Researchers are studying the rate of developing asthma. The researchers enroll 100 participants who have been determined to not have asthma. The researchers plan to follow these participants over one year to see who develops asthma, beginning on January 1st. Participants visit a doctor monthly, at the end of the month, to determine if they have asthma. After one year, 5 of the participants have developed asthma. Two participants had asthma diagnosed at the end of March. Two participants had asthma diagnosed at the end of August. One participant had asthma diagnosed at the end of November.

1) How many person-months did the study participants contribute to the study, assuming that patients became cases of asthma on the last day of the month when they were diagnosed?

2) What is the rate of asthma cases in this study?

3) In this study, when were participants removed from the denominator of the rate?

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

## Answers to Practice Questions

1)
(95 patients * 12 months)=1140
(2 patients * 3 months)=6
(2 patients*8 months)=16
(1 patient* 11 months)=11
Sum= 1140 + 6 + 16 + 11 =1173 person-months

2)
The one year rate = (# of new cases) / total person-time at risk = 5 cases / 1173 person-months = 0.0043

3)
Participants were removed when they were no longer at risk of the outcome, which was asthma. All participants began the study at-risk of developing asthma. Two patients were removed from the denominator of the rate at the end of March. Two participants were removed from the denominator of the rate at the end of August. One participant was removed from the denominator of the rate at the end of August. The remaining 95 asthma-free patients were removed from the denominator of the rate only at the very end of the study, which would have been December 31st.

# ERIC Notebook

Second Edition

## Case-Control Studies


**Second Edition Authors:**

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


Case-control studies are used to determine if there is an association between an exposure and a specific health outcome. These studies proceed from effect (e.g. health outcome, condition, disease) to cause (exposure). Case-control studies assess whether exposure is disproportionately distributed between the cases and controls, which may indicate that the exposure is a risk factor for the health outcome under study. Case-control studies are frequently used for studying rare health outcomes or diseases.

Unlike cohort or cross-sectional studies, subjects in case-control studies are selected because they have the health outcome of interest (cases). Selection is not based on exposure status. Controls, persons who are free of the health outcome under study, are randomly selected from the population out of which the cases arose. The case-control study aims to achieve the same goals (comparison of exposed and unexposed) as a cohort study but does so more efficiently, by the use of sampling.

At baseline:

- Selection of cases and controls based on health outcome or disease status
- Exposure status is unknown

After cases and controls have been identified, the investigator determines the proportion of cases and the proportion of controls that have been

*Exposure at some specified point before disease onset

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

exposed to the exposure of interest. Thus, the denominators obtained in a case-control study do not represent the total number of exposed and non-exposed persons in the source population.

After the investigator determines the exposure, a table can be formed from the study data.

| | Cases | Controls |
|---|---|---|
| Exposed | a | b |
| Unexposed | c | d |

### Measures of incidence in case-control studies

In case-control studies the proportion of cases in the entire population-at-risk is unknown, therefore one cannot measure incidence of the health outcome or disease. The controls are representative of the population-at-risk, but are only a sample of that population, therefore the denominator for a risk measure, the population- at-risk, is unknown. We decide on the number of diseased people (cases) and non-diseased people (controls) when we design our study, so the ratios of controls to cases is not biologically or substantively meaningful. However, we can obtain a valid estimate of the risk ratio or rate ratio by using the exposure odds ratio (OR).*

Odds of exposure among cases = a/c

Odds of exposure among controls = b/d

Diseased person-years

| | Disease | No Disease |
|---|---|---|
| Exposed | a | $n_1$ |
| Unexposed | c | $n_2$ |

RR = (a/n1)/(c/n2)

***Note:** Under some conditions, the odds ratio approximates a risk ratio or rate ratio. However, this is not always the case, and care should be taken to interpret odds ratios appropriately.*

Case-Control Study

| | Cases | Controls |
|---|---|---|
| Exposed | a | b |
| Unexposed | c | d |

OR = (a/c)/(b/d) = (a/b)/(c/d) = (axd)/(cxb)

If b and d (from the case-control study) are sampled from the source population, n1 + n2, then b will represent the n1 component of the cohort and d will represent the n2 component, and (a/n1)/(c/n2) will be estimated by (a/b)/(c/d).

### Interpreting the odds ratio

The odds ratio is interpreted the same way as other ratio measures (risk ratio, rate ratio, etc.).

OR = 1 Odds of disease is the same for exposed and unexposed

OR > 1 Exposure increases odds of disease

OR < 1 Exposure reduces odds of disease

For example, investigators conducted a case-control study to determine if there is an association between colon cancer and a high fat diet. Cases were all confirmed colon cancer cases in North Carolina in 2010. Controls were a sample of North Carolina residents without colon cancer. The odds ratio was 4.0. This odds ratio tells us that individuals who consumed a high fat diet have four times the odds of colon cancer than do individuals who do not consume a high-fat diet. In another study of colon cancer and coffee consumption, the OR was 0.60. Thus, the odds of colon cancer among coffee drinkers is only 0.60 times the odds among individuals who do not consume coffee. This OR tells us that coffee consumption seems to be protective against colon cancer.

### Types of case-control studies

Case-control studies can be categorized into different groups based on when the cases develop the health outcome and based on how controls are sampled. Some

case-control studies use prevalent cases while other case-control studies use incident cases. There are also different ways that cases can be identified, such as using population-based cases or hospital-based cases.

### Types of cases used in case control studies

Prevalent cases are all persons who were existing cases of the health outcome or disease during the observation period. These studies yield a prevalence odds ratio, which will be influenced by the incidence rate and survival or migration out of the prevalence pool of cases, and thus does not estimate the rate ratio. Case control studies can also use incident cases, which are persons who newly develop the health outcome or disease during the observation period. Recall that prevalence is influenced by both incidence and duration. Researchers that study causes of disease typically prefer incident cases because they are usually interested in factors that lead up to the development of disease rather than factors that affect duration.

### Selecting controls

Selection of controls is usually the most difficult part of conducting a case-control study. We will discuss 3 possible ways to select controls:

1. Base or case-base sampling
2. Cumulative density or survivor sampling
3. Incidence density or risk set sampling

### Base sampling or case-base sampling

This sampling involves using controls selected from the source population such that every person has the same chance of being included as a control. This type of sampling only works with a previously defined cohort. In these case-control studies, the odds ratio provides a valid estimate of the risk ratio without assuming that the disease is rare in the source population.

### Cumulative density sampling or survivor sampling

When controls are sampled from those people who remained free of the health outcome at the end of follow-up then we call the sampling cumulative density sampling or survivor sampling. Controls cannot ever have the outcome (become cases) when using this type of sampling. In these case-control studies, the odds ratio estimates the rate ratio only if the health outcome is rare, i.e. if the proportion of those with the health outcome among each exposure group is less than 10% (requires the rare disease assumption).

### Incidence density sampling or risk set sampling

When cases are incident cases and when controls are selected from the at-risk source population at the same time as cases occur (controls must be eligible to become a case if the health outcome develops in the control at a later time during the period of observation) then we call this type of sampling incidence density sampling or risk set sampling. The control series provides an estimate of the proportion of the total person-time for exposed and unexposed cohorts in the source population. In these case -control studies, the odds ratio estimates the rate ratio of cohort studies, without assuming that the disease is rare in the source population.

Note that it is possible, albeit rare, that a control selected at a later time point could become a case during the remaining time that the study is running. This differs from case-control studies that use cumulative density sampling or survivor sampling, which select their controls after the conclusion of the study from among those individuals remaining at risk.

Selecting controls in a risk set sampling or incidence density sampling manner provides two advantages:

1. A direct estimate of the rate ratio is possible.
2. The estimates are not biased by differential loss to follow up among the exposed vs. unexposed controls.

For example, if a large number of smokers left the source population after a certain time point, they would not be available for selection at the end of the study – when controls would be selected in a study that uses cumulative density sampling or survivor sampling. This would give the

investigators biased information regarding the level of exposure among the controls over the course of the study.

### Source populations for case-control studies

Source populations can be restricted to a population of particular interest, e.g. postmenopausal women at risk of breast cancer. This restriction makes it easier to control for extraneous confounders in the population. Controls should represent the restricted source population from which cases arise, not all non-cases in the total population. The cases in the study do not have to include all cases in the total population.

### Sources of cases

- Cases diagnosed in a hospital or clinic
- Cases entered into a disease registry, e.g. cancer, birth defects, deaths
- Cases identified through mass screening, e.g. hypertensives, diabetics
- Cases identified through a prior cohort study, e.g. lung cancers in an occupational asbestos cohort

### Sources of controls

- Population controls are non-cases sampled from the source population giving rise to cases. This is the most desirable method for selecting controls. Sampling randomly from census block groups, or a registry such as the Department of Motor Vehicles (of adults who are able to drive) are examples of ways to find and recruit population-based controls.
- Neighborhood or friend controls are appropriate for selection as controls if these individuals would be included as cases if they developed the health outcome of interest. It is not appropriate to select neighbors or friends as controls if they share the exposure of interest.
- Hospital controls - There are certain problems with hospital controls in that they may not be from the same source population from which the cases arose. Hospital controls may not be representative of the exposure prevalence in the source population of cases, e.g. there may be a higher prevalence of smokers in hospitals. Hospital controls also may have diseases resulting from the exposure of interest, e.g. the exposure (smoking) is related to the disease of interest (cancer) and to heart and lung diseases from which the controls may be suffering.
- Controls with another disease - However if the study is on lung cancer, for example, it is essential to exclude cancers known or suspected to be related to the study exposure of interest. These controls also share some of the same problems as hospital controls.

### Advantages of case-control studies

Case-control studies are the most efficient design for rare diseases and require a much smaller study sample than cohort studies. Additionally, investigators can avoid the logistical challenges of following a large sample over time. Thus, case-control studies also allow more intensive evaluation of exposures of cases and controls. Case-control studies that use incidence density sampling or risk set sampling yield a valid estimate of the rate ratio derived from a cohort study if incident cases are studied and controls are sampled from the risk set of the source population. If properly performed (i.e. appropriate sampling), case-control studies provide information that mirrors what could be learned from a cohort study, usually at considerably less cost and time.

### Disadvantages of case-control studies

Case-control studies do not yield an estimate of rate or risk, as the denominator of these measures is not defined. Case-control studies may be subject to recall bias if exposure is measured by interviews and if recall of exposure differs between cases and controls. However, investigators may be able to avoid this problem if historical records are available to assess exposure. Choosing an appropriate source population is also difficult and may contribute to selection bias. Case-control studies are not an efficient means for studying rare exposures (less than 10% of controls are exposed) because very large numbers of cases and controls are needed to detect the effects of rare exposures.

### Terminology

*Cohort studies:* An observational study in which subjects are sampled based on the presence (exposed) or absence(unexposed) of a risk factor of interest. These subjects are followed over time for the development of a health outcome of interest.

*Cross-sectional studies:* An observational study in which subjects are sampled at one point in time, and then the associations between the concurrent risk factors and health outcomes are investigated.

*Exposure odds ratio (OR):* the odds of a particular exposure among persons with a specific health outcome divided by the corresponding odds of exposure among persons without the health outcome of interest. Yields a valid estimate of the incidence rate ratio or risk ratio derived from a cohort study, depending on control sampling.

*Incident case:* a person who is newly diagnosed as a case.

*Prevalent case:* a person who has a health outcome of interest that was diagnosed in the past.

*Risk ratio (RR):* the likelihood of a particular health outcome occurrence among persons exposed to a given risk factor divided by the corresponding likelihood among unexposed persons.

*Source population:* the population out of which the cases arose.

From: Medical Epidemiology, R.S. Greenberg, 1993, 1996.

### Practice Questions

*Answers are located at the end of this notebook*

1) Researchers conduct a case-control study of breast cancer, using incident cases. The researchers find out that 90% of the cases had taken hormonal contraceptives in the past. Should the researchers conclude that hormonal contraceptives increase the risk of developing breast cancer?

2) Researchers conduct a case-control study of pancreatic cancer. The study included 200 cases and 200 controls. Of the cases, 80% reported they smoked cigarettes. Among the controls, 50% reported they smoked cigarettes.

a) Prepare a 2x2 table with these data

b) Calculate the exposure odds ratio

c) Interpret the exposure odds ratio in a sentence

### References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-.2013.

### Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. No, The information provided in this question is only for cases. No information was given in the question about the control group that was studied. To assess if exposure to hormonal contraceptives was associated with the risk of breast cancer, information would be required on the proportion of control subjects who had taken hormonal contraceptives and the researchers would need to calculate the exposure odds ratio.

2.

a) 2x2 table

| | Cases | Controls | |
|---|---|---|---|
| Exposed | 160 (a) | 100 (b) | 260 |
| Unexposed | 40 (c) | 100 (d) | 140 |
| | 200 | 200 | 400 |

b) Exposure odds ratio = (a/c) / (b/d) = (a*d)/(c*b) = (160*100) / (40*100) = 4.0

c) An odds ratio of 4.0 means that the odds of smokers being a case are 4 times the odds of non-smokers being a case.

# ERIC Notebook

Second Edition

## Causality


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

The primary goal of the epidemiologist is to identify those factors that have a causal impact on development or prevention of a health outcome, thereby providing a target for prevention and intervention. At first glance, causality may appear to be a relatively simple concept to define. However, adequately distinguishing causal agents from non-causal agents is not an easy task from an epidemiologic perspective. Unfortunately, there is no elementary parameter that can be measured to provide a definitive answer when determining causality. Rather, there are a series of criteria that have been developed and refined over the years that now serve as the guideline for causal inference. The most important point to remember is that causality is not determined by any one factor, rather it is a conclusion built on the preponderance of the evidence.

Epidemiologist Austin Bradford Hill is credited with identifying the nine factors that constitute the current standard for determining causality (1965). In his article, Hill expanded upon criteria that had previously been set forth in the report *Smoking and Health* (1964) by the United States Surgeon General. Below is a discussion of the nine criteria defined by Hill to be utilized in the determination of causality.

It is important to note that satisfying these criteria may lend support for causality, but failing to meet some criteria does not necessarily provide evidence against causality, either. Hill's causal criteria should be viewed as guidelines, not as a "checklist" that must be satisfied for a causal relationship to exist.

### Hill's causal criteria

*Strength of association*

Strength of association between the exposure of interest and the outcome is most commonly measured via risk ratios, rate ratios, or odds ratios. Hill believed that causal relationships were more likely to demonstrate strong associations than were non-causal agents. Smoking and lung cancer is a perfect example where risk ratios, rate ratios, and odds ratios are in the 20 to 40 range when comparing smokers to non-smokers. However, weak associations as demonstrated by the risk ratio, rate ratio, or odds ratio should not be taken as an indication of non-causality. This is particularly true when the outcome of interest is common.

An example of a common outcome that exhibits a weak association to smoking is cardiovascular disease (CVD). Yet even with a weak association, evidence supports the casual nature between smoking and the development of CVD. Further-more, one should not assume that a strong association alone is indicative of causality, as the presence of strong confounding may erroneously lead to a strong causal association.

*Consistency of data*

This tenant refers to the reproducibility of results in various populations and situations.

Consistency is generally utilized to rule out other explanations for the development of a given outcome. It should also be noted that a lack of consistency does not negate a causal association as some causal agents are causal only in the presence of other co-factors. In general, the greater the consistency, the more likely a causal association.

*Specificity*

This criterion has been proven to be invalid in a number of instances, with smoking being the primary example. Evidence clearly demonstrates that smoking does not lead solely to lung carcinogenesis but to a myriad of other clinical disorders ranging from emphysema to heart disease. On the other hand, there are certain situations where a 1 to 1 relationship exists, such as with certain pathogens which are necessary to produce a specific disease. Tuberculosis is a good example.

*Temporality*

This criterion has been identified as being the most likely to be the sine qua non for causality, i.e. it is absolutely essential. For an agent to be causal, its presence must precede the development of the outcome. Lack of temporality rules out causality. An example found in the literature is the relationship between atrial fibrillation (AF) and pulmonary embolism. Current wisdom supports that pulmonary embolism causes atrial fibrillation, however more recent evidence and plausible biological hypothesis suggest that the reverse could be true. Determining the proper course of care may hinge upon discovering if pulmonary emboli can indeed precede and thus perhaps cause the development of atrial fibrillation.

*Dose-response*

The presence of a dose-response relationship between an exposure and outcome provides good evidence for a causal relationship; however, its absence should not be taken as evidence against such a relationship. Some diseases or health outcomes do not display a dose-response relationship with a causal exposure. They may demonstrate a threshold association where a given level of exposure is required for disease or health outcome initiation, and any additional exposure does not affect the outcome.

*Biological plausibility*

Support for this criterion is generally garnered in the basic science laboratory. It is not unusual for epidemiological conclusions to be reached in the absence of evidence from the laboratory, particularly in situations where the epidemiological results are the first evidence of a relationship between an exposure and an outcome. However, one can further support a causal relationship with the addition of a reasonable biological mode of action, even though basic science data may not yet be available.

*Coherence*

This term represents the idea that, for a causal association to be supported, any new data should not be in opposition to the current evidence, that is, providing evidence against causality. However, one should be cautious in making definite conclusions regarding causation, since it is possible that conflicting information is incorrect or highly biased.

*Experimental evidence*

Today's understanding of Hill's criteria of experimental evidence results from many areas: the laboratory, epidemiological studies, and preventive and clinical trials. Ideally, epidemiologists would like experimental evidence obtained from a well-controlled study, specifically randomized trials. These types of studies can support causality by demonstrating that "altering the cause alters the effect".

*Analogy*

This is perhaps one of the weaker of the criteria in that analogy is speculative in nature and is dependent upon the subjective opinion of the researcher. An example of an analogy is that while infection may cause a fever, not all fevers are due to infection. Absence of analogies should not be taken as evidence against causation.

*Other considerations*

In addition to assessing the components of Hill's list, it is also critically important to have a thorough understanding of the literature to determine if any other plausible explanations have been considered and tested previously.

## Additional models for causality

In additional to Hill's guidelines for causality, several recent models for understanding causality have been developed. These include Kenneth Rothman's component cause theory, counterfactual models, and directed acyclic graphs. We provide brief descriptions below but refer you to the suggested readings section for more information.

*Component causes*

Kenneth Rothman described the circumstances leading to a health outcome as being parts of one pie chart, or a "causal pie." Without each component in place, the disease or health outcome would not have occurred at that specific point in time.

**Example**

If you were to slip on an icy sidewalk and break your wrist, there may be a number of factors that contributed to that outcome. First, the fall directly leads to a broken wrist, and this was a necessary component of the outcome. However, which factors led to you walking on the ice and having a fall in the first place, and what other factors influenced your broken wrist? Maybe you were wearing poor footwear and were tugged by your dog who was chasing a squirrel. Maybe you didn't receive enough calcium in your diet, developed osteoporosis, and the fall would not have broken your wrist without weakened bones. Each of these components ultimately led to the outcome.

**Figure: Causal Pie**

Rothman's component causes theory is one way to consider all factors involved in the development of a health outcome.

*Counterfactual models*

Many statistical models that are used to adjust for confounding are based on counterfactual thinking. These models are based on comparing an exposed group of people to a fictional group of people who are exactly the same except they are unexposed to the key variable. These models try to answer the question: "If this one experience or exposure in the past did not happen to an individual, how would it impact that person's health outcome today?" Of course, this is an impossible situation; we cannot go back in time and change an individual's exposure status and track both outcomes over time. However, using two very similar populations of people – one group that is exposed and another that is unexposed – it is possible to estimate that very counterfactual situation on a group level. Many statistical modeling programs that adjust for potential confounders are modeling a counterfactual scenario to produce a less biased measure of association.

*Directed acyclic graphs (DAGs)*

DAGs are one method used to create a conceptual diagram that maps the relationships between the main exposure, outcome of interest, and all potential confounders for a given study. Through a DAG analysis, specific rules are followed to determine if confounding might be present for a given research question. Once confounders have been identified and adjusted for, a less biased measure of association can be obtained.

**Figure: Basic Directed acyclic graph (DAG)**

**Practice Questions**

*Answers are located at end of this notebook.*

1) Researchers conducted a cohort study of the association between air pollution and asthma. The rate ratio was 8.0, when comparing those exposed to high levels of air pollution with those exposed to low levels of air pollution. Which of the following issues should the researchers consider when making their study conclusions and when thinking about causality? Choose all that apply.

a) The rate ratio of 8.0 indicates a strong association, which lends support for causality
b) Strong confounding may actually be causing the strong association seen
c) Other studies of the same exposure—health outcome association reported rate ratios in the range of 1.5- 3.0, less than the rate ratio of 8.0 seen in this study
d) The temporal sequence of the exposure and outcome should be known in order to draw accurate conclusions

2) Considering the component causes model of causality developed by Kenneth Rothman, which of the following factors may play a role in the cause of a car accident? Choose all that apply.

a) The road conditions
b) The condition of the car
c) The amount of sleep that the driver of the car got the previous night
d) The level of distraction of the driver
e) The driver's overall health

### References and Suggested Readings

Flegel KM. When atrial fibrillation occurs with pulmonary embolism, is it the chicken or the egg? *CMAJ*. 160 (8):1181-2, 1999.

Greenland S, Holland PW, Mantel N, Wickramaratne PJ and Holford TR. Confounding in Epidemiologic Studies. *Biometrics*. 45(4):1309-1322

Hill AB. The environment and disease: association or causation? *Proc R Soc Med.* 58:295-300, 1965.

Mengersen KL. Merrilees MJ. Tweedie RL. Environmental tobacco smoke and ischaemic heart disease: a case study in applying causal criteria. *Int Arch of Occup & Env Health*. 72 Suppl:R1-40, 1999.

Ridgway D. The logic of causation and the risk of paralytic poliomyelitis for an American child. *Epidemiology & Infection*. 124(1):113-20, 2000.

Rothman KJ. Causes. *Am J Epidemiol* 104:587–92, 1976.

Rothman KJ, Greenland S, Poole C, and Lash TL. Causation and causal inference. In: *Modern Epidemiology (3e)*. Edited by: Rothman KJ, Greenland S, and Lash TL. Philadelphia: Lippencott-Raven Publishers; 2008:5-31.

Shrier I and Platt RW. Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology* 2008, 8:70.

United States Department of Health, Education, and Welfare. *Smoking and health:* Report of the Advisory Committee to the Surgeon General of the Public Health Service. Washington, D.C Government Printing Office, 1964 PHS Publ. No. 1103.

Weed DL. On the use of causal criteria. *Int J of Epidemiology*. 26(6):1137-41, 1997.

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

### Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Sandra Demming, MPH, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

### Answers to Practice Questions

1) All answer choices are correct. When making their study conclusions and considering causality, the researchers would need to consider all of the issues listed here as well as additional issues. It is true that a rate ratio of 8.0 indicates a strong association. Such a large rate ratio would lend support for causality if this rate ratio is indeed correct. However, if strong confounding is present in a study then a strong association may actually be seen in error so the researchers would need to consider the presence of and extent of confounding. It would also be important for the researchers to consider previous studies of the same exposure—health outcome association. If all previous literature had reported rate ratios in the range of only 1.5- 3.0, then the researchers would want to be especially careful in determining if their results were really accurate and non-biased. Finally, the timing of the exposure would also be important to consider. Temporality is a key criterion because, for a relationship to be causal, the exposure must precede the outcome. The researchers would have needed to carefully set up their study to ensure that they were studying participants who were first exposed to air pollution and then went on to develop asthma, rather than studying participants who already had asthma prior to being exposed to high levels of air pollution.

2) All answer are correct. Kenneth Rothman's component cause theory is another way to understand causality. It is possible that a particular car accident may have occurred only because all of these factors occurred. Perhaps if even one of these factors had not occurred, the car accident would not have occurred. For example, this

fictitious car accident may have occurred on a day with wet road conditions. The car's tires may have not been in good condition, leading to a further decrease in stopping time on wet roads. The driver may have stayed up late the previous night, leading to a slower reaction time due to not getting sufficient sleep. The driver's reaction time might still have been quick enough to avoid the accident, however, due to holding a cup of coffee (a distraction) and due to having arthritis (a health problem), the driver was unable to react in time to steer the car away from a collision. Furthermore, chance may also have played a role in the accident. It is possible that had the driver not been on a particular portion of the road, at a particular time, that the accident would not have happened. All of these factors may have contributed in unison to lead to the occurrence of the car accident.

# ERIC Notebook

Second Edition

## Cohort Studies


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

A cohort study is a type of epidemiological study in which a group of people with a common characteristic is followed over time to find how many reach a certain health outcome of interest (disease, condition, event, death, or a change in health status or behavior). A cohort is defined as a group of persons, usually 100 or more in size, who share a common characteristic, e.g. smokers, workers in a lead smelter, people born in the same year, or all enrollees of a specific health insurance plan Cohort studies compare an exposed group of individuals to an unexposed (or less exposed) group of individuals to determine if the outcome of interest is associated with exposure. There are two types of cohort studies: prospective and retrospective (or historical) cohorts. Prospective studies follow a cohort into the future for a health outcome, while retrospective studies trace the cohort back in time for exposure information after the outcome has occurred. Both types of cohort studies are also referred to as longitudinal or follow-up studies.

### Establishing the cohort

The investigator controls the selection of the cohort. The investigator may choose a cohort based on age, location, exposure to a certain working environment, or some other common characteristic. Cohorts may be selected on the basis of exposures known at baseline, e.g. smokers vs. nonsmokers. Alternatively, cohorts may be divided into exposure categories once baseline measurements of a defined population are made. For example, the Framingham Cardiovascular Disease Study (CVD) used baseline measurements to divide the population into categories of CVD risk factors.

For instance, an investigator wants to study whether exposure to military aircraft engine noise is a risk factor for hearing loss. The cohort this investigator would want to establish should be composed of two groups of military personnel: one exposed to engine aircraft noise (the group under study) and the other unexposed to engine aircraft noise (a comparison group). The unexposed group should be representative of the exposed group on all factors except exposure.

### The cohort at baseline

After the cohort of study subjects is established, their individual exposures of interest are identified at baseline (through interviews, questionnaires, bioassays, medical records, etc.). Subjects with the outcome of interest at baseline are excluded. Therefore, all members of the cohort are at risk of developing the outcome at the beginning of observation.

Following the last example, anyone in the cohort of military personnel with a specified hearing loss at baseline would be excluded from the cohort and would not be followed.

### Following the cohort

The cohort is then followed over time for new occurrences of the outcome of

interest, in the above example, hearing loss. In a prospective, or concurrent, cohort study baseline exposure is assessed at the beginning of the study and the cohort is followed into the future. In a retrospective, or historical cohort study, baseline exposure is assessed at some point in the past through historical records, e.g. health records for a cohort of factory workers may provide exposure and outcome information up to the present.

| Cohort | Baseline | Follow-up |
|---|---|---|
| Prospective | Assessed at beginning of study | Followed into the future for outcome |
| Retrospective | Assessed at some point in the past via historical records | Outcome has already occurred and is assessed via historical records |

Cohorts are followed over time to the end of follow-up. Occurrence of the outcome of interest may be determined via interviews with members of the cohort and/or family members, or by viewing health and/or work records to conclude the study.

**The basic design of a cohort study from beginning of the study to end of follow-up. E = exposed, Ē = not exposed, D = diseased and D̄ = not diseased.**

**Evaluation of the results**

During the follow-up period the investigator counts the number of subjects who develop the outcome of interest. This count is the numerator for a calculation of risk, also referred to as **cumulative incidence** or **incidence proportion**. The number of persons at risk at baseline is the denominator.

*Risk (also called Cumulative Incidence or Incidence Proportion) = new occurrences of the outcome / population-at-risk at baseline*

Two risks can be compared to provide a risk ratio. The reference group is a comparable unexposed cohort. The index group is the exposed cohort. The risk ratio is computed by dividing the risk in the exposed group by the risk in the unexposed group. The risk ratio gives a relative measure of the increase or decrease in incidence between the exposed and unexposed groups.

$$\text{Risk Ratio} = Risk_{exposed}/Risk_{unexposed}$$

As with risk, **an incidence rate measure (IR)** is calculated with new occurrences of the outcome as the numerator. An incidence rate is also called incidence density. However, in an IR calculation the denominator is person-time (days, months, or years) at risk during follow-up. Person-time is measured by summing the total time each member of the cohort was free of the outcome of interest and thus contributed to person-time-at-risk during the follow-up period. The IR measures the rapidity of occurrence of new health outcomes in the population.

$$\text{Incidence rate} = \frac{new\ occurrences\ of\ the\ outcome}{person\text{-}time\ at\ risk}$$

Two IRs may also be compared to find the relative increase or decrease in the rate of health outcome occurrence between the exposed and unexposed groups. This relative measure is called the **incidence rate** ratio (IRR) or the rate ratio.

$$\text{Incidence rate ratio} = \frac{IR_{exposed}}{IR_{unexposed}}$$

Incidence measures between exposed and unexposed cohorts can also be subtracted from one another to find the difference between the two

measures. This measure is referred to as a **risk difference** or a **rate difference**.

$$\text{Risk Difference} = Risk_{exposed} - Risk_{unexposed}$$

$$\text{Rate Difference} = I_{exposed} - I_{unexposed}$$

Exposure may be a risk factor or a preventive factor in the development of the outcome of interest. When exposure is preventive, the risk ratio or rate ratio will be less than one.

### Advantages of a cohort study

A cohort study can be used to directly measure the risk and rate of a health outcome occurrence over time. Cohort studies are an efficient means of studying rare exposures (e.g. gasoline fumes, as discussed in the next paragraph), in contrast to case-control studies, which tend to be better for rare outcomes. Cohort studies also allow the investigator to assess multiple outcomes of a single exposure.

> A cohort study would be the most efficient means of studying the effects of long-term exposure to gasoline fumes. The cohort would consist of individuals who are exposed daily to gasoline fumes (auto mechanics, gas station attendants, sea crewman on tankers, etc.). By studying this group
> of individuals, the investigator can better determine the direct effects of long-term, regular gasoline inhalation. Also, by conducting a cohort study, an investigator could determine if gasoline inhalation causes many different health outcomes (e.g., different types of cancer and respiratory illnesses).

### Additional advantages of cohort studies

Cohort studies establish temporal relationships between exposure and outcome. Exposure clearly precedes the outcome because the population under study at baseline is free of the outcome of interest. Cohort studies also avoid recall bias (as the exposure is determined before the outcome, one's health outcome or disease state won't affect how accurately one recalls exposure levels), as well as, survival bias (duration of disease influencing exposure measurements). Therefore, cohort studies are the best observational study design used to help establish cause and effect relationships.

### Disadvantages of cohort studies

Cohort studies often require large sample sizes, especially when the outcome is rare, defined as less than 1 event per 1000 person- years (e.g., all specific cancers). Therefore, cohort studies tend to be expensive and time-consuming. When there are *losses to follow-up* (individuals who leave the cohort before the end of follow-up) biases may occur. Thus, individuals who leave the cohort prematurely may have a different baseline risk than the members who remain in the cohort throughout the entire length of follow-up. Therefore, the study may not be *generalizable* to the original target population, but only to those who remained under investigation throughout the length of the study. Also, any differences in the quality of measurement of exposure or disease between exposed and non-exposed cohorts may introduce *information bias* and thereby distort the results.

### Practice Questions

*Answers are located at end of this notebook.*

Researchers are conducting a prospective cohort study of the association between being an office worker who uses a computer daily and carpal tunnel syndrome (a hand/ arm nerve condition).

1) Researchers choose a group of office workers from one company to be their "exposed" group. Which of the following "unexposed or less exposed" comparison groups would be the <u>best</u> choice?

a) A group of high school students, who use computers daily

b) Another group of office workers at a different company, who do not use computers daily

c) Another group of office workers at the same company, who do not use computers daily

d) A group of professional golfers, who use their hands/ arms intensively for their sport

2) Who should be excluded from the cohort (not included in the study)?

3) A total of 300 exposed and 300 unexposed participants are enrolled and followed for 10 years. A total of 25 exposed and 17 unexposed had the outcome of interest over the follow-up period.

a) What is the risk of developing carpal tunnel syndrome among the exposed?

b) What is the risk of developing carpal tunnel syndrome among the unexposed?

c) What is the risk ratio for developing carpal tunnel syndrome?

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

## Answers to Practice Questions

1) The best answer choice is answer c, another group of office workers at the same company, who do not use computers daily. The unexposed group should be representative of the exposed group on all factors except exposure. If the researchers used students as a comparison group, they would very likely be studying a completely different age-group of participants who would likely be quite different from a population of office workers. Similarly, if the researchers used the golfers as a comparison group, they would likely be studying a group of participants that vastly differed from office workers. Studying another group of office workers from a different company would be a fairly good choice since then "unexposed" office workers would be compared with the "exposed" office workers. However, the best choice would be to compare office workers at the same company, who live in the same area, looking at those who used a computer daily versus those who did not.

2) The researchers should exclude any participants who already have carpal tunnel syndrome at the start of the study.

3a) The 10-year risk is 25/300=0.083
3b) The 10-year risk is 17/300=0.057
3c) Risk exposed / Risk unexposed = 0.083 / 0.057 = 1.47 The risk of developing carpal tunnel syndrome among office workers exposed to daily computer use is 1.47 times the risk among office workers who do not use a computer daily.

# ERIC Notebook

Second Edition

## Common Measures and Statistics in Epidemiological Literature


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


For the non-epidemiologist or non-statistician, understanding the statistical nomenclature presented in journal articles can sometimes be challenging, particularly since multiple terms are often used interchangeably, and still others are presented without definition. This notebook will provide a basic introduction to the terminology commonly found in epidemiological literature.

### Measures of frequency

Measures of frequency characterize the occurrence of health outcomes, disease, or death in a population. These measures are descriptive in nature and indicate how likely one is to develop a health outcome in a specified population. The three most common measures of health outcome or frequency are risk, rate, and prevalence.

*Risk*

Risk, also known as incidence, cumulative incidence, incidence proportion, or attack rate (although not really a rate at all) is a measure of the probability of an unaffected individual developing a specified health outcome over a given period of time. For a given period of time (i.e.: 1 month, 5 years, lifetime):

A 5-year risk of 0.10 indicates that

$$\text{Risk} = \frac{\#\ of\ new\ cases}{total\ \#\ individuals\ at\ risk}$$

an individual at risk has a 10% chance of developing the given health outcome over a 5-year period of time.

Risk is generally measured in prospective studies as the population at risk can be defined at the start of the study and followed for the development of the health outcome. However, risk cannot be measured directly in case-control studies as the total population at risk cannot be defined. Thus, in case-control studies, a group of individuals that have the health outcome and a group of individuals that do not have the health outcome are selected, and the odds of developing the health outcome are calculated as opposed to calculating risk.

$$\text{Odds} = \frac{\#\ individuals\ with\ the\ health\ outcome}{\#\ individuals\ without\ the\ health\ outcome}$$

*Rate*

A rate, also known as an incidence rate or incidence density, is a measure of how quickly the health outcome is occurring in a population. The numerator is the same as in risk, but the denominator includes a measure of person-time, typically person-years. (Person-time is defined as the sum of time that each at-risk individual contributes to the study).

$$\text{Rate} = \frac{\#\ of\ new\ cases}{total\ person\text{-}time\ at\ risk}$$

Thus a rate of 0.1 case/person-years indicates that, on average, for every 10 person-years (i.e.: 10 people each followed 1 year or 2 people followed

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

for 5 years, etc.) contributed, 1 new case of the health outcome will develop.

*Prevalence*

Prevalence is the proportion of a population who has the health outcome at a given period of time. Prevalence is generally the preferred measure when it is difficult to define onset of the health outcome or disease (such as asthma), or any disease of long duration (e.g. chronic conditions such as arthritis). A limitation of the prevalence measure is that it tends to favor the inclusion of chronic diseases over acute ones. Also, inferring causality is troublesome with prevalence data, as typically both the exposure and outcome are measured at the same time. Thus it may be difficult to determine if the suspected cause precedes the outcome of interest.

$$\text{Prevalence} = \frac{\#\ of\ affected\ individuals}{total\ \#\ of\ individuals\ in\ the\ population}$$

Thus a population with a heart disease prevalence of 0.25 indicates that 25% of the population is affected by heart disease at a specified moment in time.

A final note, risk and rates can also refer to deaths in a population and are termed mortality and mortality rate, respectively.

**Measures of association**

Measures of association are utilized to compare the association between a specific exposure and health outcome, They can also be used to compare two or more populations, typically those with differing exposure or health outcome status, to identify factors with possible etiological roles in health outcome onset. Note that evidence of an association does not imply that the relationship is causal; the association may be artifactual or non-causal as well. Common measures of association include the risk difference, risk ratio, rate ratio and odds ratio.

*Risk difference*

Risk difference is defined as

$$= Risk_{exposed} - Risk_{unexposed}$$

$$\frac{\#\ cases\ in\ exposed\ group}{total\ \#\ at\ risk\ in\ exposed\ group} - \frac{\#\ cases\ in\ control\ group}{total\ \#\ at\ risk\ in\ control\ group}$$

The risk difference, also know as the attributable risk, provides the difference in risk between two groups indicating how much excess risk is due to the exposure of interest. A positive risk difference indicates excess risk due to the exposure, while a negative result indicate that the exposure of interest has a protective effect against the outcome. (Vaccinations would be a good example of an exposure with a protective effect). This measure if often utilized to determine how much risk can be prevented by an effective intervention.

*Risk ratio and rate ratio*

Risk ratios or rate ratios are commonly found in cohort studies and are defined as: the ratio of the risk in the exposed group to the risk in the unexposed group or the ratio of the rate in the exposed group to the rate in the unexposed group

$$\text{Risk Ratio} = Risk_{exposed}/Risk_{unexposed}$$

$$\text{Rate Ratio} = Rate_{exposed}/Rate_{unexposed}$$

Risk ratios and rate ratios are measures of the strength of the association between the exposure and the outcome. How is a risk ratio or rate ratio interpreted? A risk ratio of 1.0 indicates there is no difference in risk between the exposed and unexposed group. A risk ratio greater than 1.0 indicates a positive association, or increased risk for developing the health outcome in the exposed group. A risk ratio of 1.5 indicates that the exposed group has 1.5 times the risk of having the outcome as compared to the unexposed group. Rate ratios can be interpreted the same way but apply to rates rather than risks.

A risk ratio or rate ratio of less than 1.0 indicates a negative association between the exposure and outcome in the exposed group compared to the unexposed group. In this case, the exposure provides a protective effect. For example, a rate ratio of 0.80 where the exposed group received a vaccination for Human Papillomavirus (HPV) indicates that the exposed group (those who received the vaccine) had 0.80 times the rate of HPV compared to those who were unexposed (did not receive the vaccine).

One of the benefits the measure risk difference has over the risk ratio is that it provides the absolute difference in risk, information that is not provided by the ratio of the two. A risk ratio of 2.0 can imply both a doubling of a very small or large risk, and one cannot determine which is the case unless the individual risks are presented.

*Odds ratio*

Another measure of association is the odds ratio (OR). The formula for the OR is:

$$\text{Odds ratio} = \frac{odds_{exposed}}{odds_{unexposed}}$$

The odds ratio is used in place of the risk ratio or rate ratio in case-control studies. In this type of study, the underlying population at risk for developing the health outcome or disease cannot be determined because individuals are selected as either diseased or non-diseased or as having the health outcome or not having the health outcome. An odds ratio may approximate the risk ratio or rate ratio in instances where the health outcome prevalence is low (less that 10%) and specific sampling techniques are utilized, otherwise there is a tendency for the OR to overestimate the risk ratio or rate ratio.

The odds ratio is interpreted in the same manner as the risk ratio or rate ratio with an OR of 1.0 indicating no association, an OR greater than 1.0 indicating a positive association, and an OR less than 1.0 indicating a negative, or protective association.

*The null value*

The null value is a number corresponding to no effect, that is, no association between exposure and the health outcome. In epidemiology, the null value for a risk ratio or rate ratio is 1.0, and it is also 1.0 for odds ratios and prevalence ratios (terms you will come across). A risk ratio, rate ratio, odds ratio or prevalence ratio of 1.0 is obtained when, for a risk ratio for example, the risk of disease among the exposed is equal to the risk of disease among the unexposed.

Statistical testing focuses on the *null hypothesis*, which is a statement predicting that there will be no association between exposure and the health outcome (or between the assumed cause and its effect), i.e. that the risk ratio, rate ratio or odds ratio will equal 1.0. If the data obtained from a study provide evidence against the null hypothesis, then this hypothesis can be rejected, and an alternative hypothesis becomes more probable.

For example, a null hypothesis would say that there is no association between children having cigarette smoking mothers and the incidence of asthma in those children. If a study showed that there was a greater incidence of asthma among such children (compared with children of nonsmoking mothers), and that the risk ratio of asthma among children of smoking mothers was 2.5 with a 95% confidence interval of 1.7 to 4.0, we would reject the null hypothesis. The alternative hypothesis could be expressed in two ways: 1) children of smoking mothers will have either a higher or lower incidence of asthma than other children, or 2) children of smoking mothers will only have a higher incidence of asthma. The first alternative hypothesis involves what is called a *"two-sided test"* and is used when we simply have no basis for predicting in which direction from the null value exposure is likely to be associated with the health outcome, or, in other words, whether exposure is likely to be beneficial or harmful. The second alternative hypothesis involves a *"one-sided test"* and is used when we have a reasonable basis to assume that exposure will only be harmful (or if we were studying a therapeutic agent, that it would only be beneficial).

**Measures of significance**

*The p-value*

The "p" value is an expression of the probability that the difference between the observed value and the null value has occurred by "chance", or more precisely, has occurred simply because of sampling variability. The smaller the "p" value, the less likely the probability that sampling variability accounts for the difference. Typically, a "p" value less than 0.05, is used as the decision point, meaning that there is less than a 5% probability that the difference between the observed risk ratio, rate ratio, or odds ratio and 1.0 is due to sampling variability. If the "p" value is less than 0.05, the observed risk ratio, rate ratio, or odds ratio is often said to be "statistically significant." However, the use of 0.05 as a cut-point is arbitrary. The exclusive use of "p" values for interpreting results of epidemiologic studies has been strongly discouraged in the more recent texts and literature because research on human health is not conducted to reach a decision point (a "go" or "no go" decision), but rather to obtain evidence that there is reason for concern about certain exposures or lifestyle practices or other factors that may adversely influence the health of the public. Statistical tests of significance, (such as p-values) were developed for industrial quality-control purposes, in order to make a decision whether the manufacture of some item is achieving acceptable quality. We are not making such decisions when we interpret the results of research on human health.

The lower bound of the 95% confidence interval is also often utilized to decide whether a point estimate is statistically significant, i.e. whether the measure of effect (e.g. the ratio 2.5 with a lower bound of 1.8) is statistically different than the null value of 1.0.

## Measures of precision

*Confidence interval*

A *confidence interval* expresses the extent of potential variation in a point estimate (the mean value or risk ratio, rate ratio, or odds ratio). This variation is attributable to the fact that our point estimate of the mean or risk ratio, rate ratio, or odds ratio is based on some sample of the population rather than on the entire population.

For example, from a clinical trial, we might conclude that a new treatment for high blood pressure is 2.5 times as effective as the standard treatment, with a 95% confidence interval of 1.8 to 3.5. 2.5 is the *point estimate* we obtain from this clinical trial. But not all subjects with high blood pressure can be included in any study, thus the estimate of effectiveness, 2.5, is based on a particular sample of people with high blood pressure. If we assume that we could draw other samples of persons from the same underlying population as the one from which subjects were obtained for this study, we would obtain a set of point estimates, not all of which would be exactly 2.5. Some samples would be likely to show an effectiveness less than 2.5, and some greater than 2.5.

The 95% CI is an interval that will contain the true, real (population) parameter value 95% of the time if you repeated the experiment/study. So if we were to repeat the experiment/study, 95 out of 100 intervals would give an interval that contains the true risk ratio, rate ratio or odds ratio value. Remember, that you can only interpret the CI in relation to talking about repeated sampling. Thus we can also say that the new treatment for high blood pressure is 2.5 times as effective as the standard treatment, but this measure could range from a low of 1.8 to a high of 3.5.

The confidence interval also provides information about how precise an estimate is. The tighter, or narrower, the confidence interval, the more precise the estimate. Typically, larger sample sizes will provide a more precise estimate. Estimates with wide confidence intervals should be interpreted with caution.

## Other terms

*Crude and adjusted values*

There are often two types of estimates presented in research articles, *crude* and *adjusted* values. Crude estimates refer to simple measures that do not account for other factors that may be driving the estimate. For instance, a crude death rate would simply be the number of deaths in a calendar year divided by the average population for that year. This may be an appropriate measure in certain circumstances but could become problematic if you want to compare two or more populations that vary on specific factors known to contribute to the death rate. For example, you may want to compare the death rate for two populations, one of which is located in a high air pollution area, to determine if air pollution levels affect the death rate. The high air pollution population may have a higher death rate, but you also determine that it is a much older population. As older individuals are more likely to die, age may be driving the death rate rather than the pollution level. To account for the difference in age distribution of the populations, one would want to calculate an *adjusted* death rate that adjusts for the age structure of the two groups. This would remove the effect of age from the effect of air pollution on mortality.

Adjusted estimates are a means of controlling for confounders or accounting for effect modifiers in analyses. Some factors that are commonly adjusted for include gender, race, socioeconomic status, smoking status, and family history.

## Practice Questions

*Answers are at the end of this notebook.*

1. Based on the following table, calculate the requested measures. Also provide the definition for each measure in one sentence.

a) The risk ratio comparing the exposed and the unexposed study participants

b) The risk difference between the exposed and the unexposed study participants

c) The prevalence of the disease among the entire study sample, assuming the disease is a long-term, chronic disease with no cure and assuming no study participants have died.

| | Has disease | Does not have disease | Total |
|---|---|---|---|
| **Exposed** | 651 | 450 | 1101 |
| **Unexposed** | 367 | 145 | 512 |
| **Total** | 1018 | 595 | 1613 |

2. Interpret the following risk ratios in words.

a) A risk ratio= 1.0 in a study where researchers examined the association between consuming a certain herbal supplement (the exposure) and developing arthritis.
b) A risk ratio= 2.6 in a study where researchers examined the association between ever having texted while driving (the exposure) and being in a car accident.
c) A risk ratio = 0.75 in a study where researchers examined the association between ≥ 30 minutes of daily exercise (the exposure) and heart disease.

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Answers to Practice Questions

1.a) Risk ratio= risk exposed / risk unexposed = (651/1101 ) / (367/ 512) = 0.82

The risk ratio reflects the ratio of the risk of the disease in the exposed study participants compared with the risk of the disease in the unexposed study participants.

1b) Risk difference = risk exposed - risk unexposed = (651/1101 ) - (367/512 ) = -0.13

The risk difference indicates how much excess risk is due to the exposure studied.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

## Answers Continued

1c) Prevalence= Total # people with the disease / total # of people in the study population = 1018/1613 = 0.63 Prevalence refers to the proportion of the population studied that has the disease at a given time.

2a) A risk ratio of 1.0 means there is no difference in risk for the health outcome when comparing the exposed and unexposed groups, i.e. the herbal supplement was not associated in any way with the development of arthritis

2b) A risk ratio of 2.6 means there is a positive association, i.e. there is an increased risk for the health outcome among the exposed group when compared with the unexposed group. The exposed group has 2.6 times the risk of having the health outcome when compared with the unexposed group. In this example, the risk ratio of 2.6 means that people who had reported ever texting while driving had 2.6 times the risk of being in a car accident when compared with people who reported never having texted while driving.

2c) A risk ratio of 0.75 means there is an inverse association, i.e. there is a decreased risk for the health outcome among the exposed group when compared with the unexposed group. The exposed group has 0.75 times the risk of having the health outcome when compared with the unexposed group. In this example, the risk ratio of 0.75 means that people who exercised at least 30 minutes per day had 0.75 times the risk of developing heart disease when compared with people who did not exercise at least 30 minutes a day.

# ERIC Notebook

Second Edition

## Common Statistical Tests and Applications in Epidemiological Literature


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Any individual in the medical field will, at some point, encounter instances when epidemiological methods and statistics will be valuable tools in addressing research questions of interest.

Examples of such questions might include:

- Will treatment with a new anti-hypertensive drug significantly lower mean systolic blood pressure?
- Is a visit with a social worker, in addition to regular medical visits, associated with greater satisfaction of care for cancer patients as compared to those who only have regular medical visits?

There are a number of steps in evaluating data before actually addressing the above questions. These steps include description of your data as well as determining what the appropriate tests are for your data.

### Description of data

The type of data one has determines the statistical procedures that are utilized. Data are typically described in a number of ways: by type, distribution, location and variation.

There are three different types of data: nominal, ordinal, and continuous data. Nominal data do not have an established order or rank and contain a finite number of values. Gender and race are examples of nominal data. Ordinal data have a limited number of values between which no other possible values exist. Number of children and stage of disease are good examples of ordinal data. It should be noted that ordinal data do not have to have evenly spaced values as occurs with continuous data, however, there is an implied underlying order. Since both ordinal and nominal data have a finite number of possible values, they are also referred to as discrete data. The last type of data is continuous data which are characterized by having an infinite number of evenly spaced values. Blood pressure and age fall into this category. It should be noted for data collection and analysis that continuous, ordinal, or nominal values can be grouped. Grouped data are often referred to as categorical. Possible categories might include: low, medium, high, or those representing a numerical range.

A second characteristic of data description, distribution, refers to the frequencies or probabilities with which values occur within our population. Discrete data are often represented graphically with bar graphs like the one below (Figure 1).

Figure 1. Bar graph

Continuous data are commonly assumed to have a symmetric, bell-shaped curve as shown below (Figure 2). This is known as a Gaussian distribution, the most commonly assumed distribution in statistical analysis.

Figure 2. Gaussian distribution

### Hypothesis testing

Hypothesis testing, also known as statistical inference or significance testing, involves testing a specified hypothesized condition for a population's parameter. This condition is best described as the null hypothesis. For example, in a clinical trial of a new anti-hypertensive drug, the null hypothesis would state that there is no difference in effect when comparing the new drug to the current standard treatment. Contrary to the null is the alternative hypothesis, which generally defines the possible values for a parameter of interest. For the previous example, the alternative hypothesis is that there is a difference in the mean blood pressure of the standard treatment and new drug group following therapy. The alternative hypothesis might also be described as your "best guess" as to what the values are.

However, in statistical analysis, the null hypothesis is the main interest, and is the one actually being tested. In statistical testing, we assume that the null hypothesis is correct and determine how likely we are to have obtained the sample (or values) we actually obtained in our study under the condition of the null. If we determine that the probability of obtaining the sample we observed is sufficiently small, then we can reject the null hypothesis. Since we are able to reject the null hypothesis, we have evidence that the alternative hypothesis may be true.

On the other hand, if the probability of obtaining our study results is not small, we fail to reject the assumption that the null hypothesis is true. It should be noted that we are not concluding that the null is true. This is a small, but important distinction. A test that fails to reject the null hypothesis should be considered inconclusive. An example will help to illustrate this point.

In a sealed bag, we have 100 blue marbles and 20 red marbles. (This bag is essentially representing the entire population). One individual formulates the null hypothesis that "all the marbles are blue", and the alternative which is "all the marbles are not blue". To test this hypothesis, 10 marbles are sampled from the bag. All ten marbles selected are indeed blue. Thus the individual has failed to reject the null that all the marbles in the bag are blue. However, because all of the marbles were not sampled, you cannot conclude that all the marbles in the bag are blue. (We happen to know this is not true, but it is impossible to know in the real world with populations too large to fully evaluate). If another individual selects 10 marbles from the bag and finds that 8 are blue and 2 are red, we can reject the null hypothesis that all the marbles are blue since we have selected at least one red marble.

### Error in statistical testing

Earlier, we indicated that we can reject the null hypothesis if the probability of obtaining a sample like the one observed in our study is sufficiently small. You may ask "What is sufficiently small?" "How small" is determined by how willing we are to reject the null hypothesis when it accurately reflects the population from which it is sampled. This type of error is called a *Type I error*. This error is also commonly called alpha (α). Alpha is the probability of rejecting the null hypothesis when the null is true. This probability is selected by the researcher and is typically set at 0.05. It is important to remember that this is an arbitrary cut-point and should be taken into consideration when making conclusions about the results of the study.

There is a second type of error that can be made during statistical testing. It is known as *Type II error,* which is the probability of not rejecting the null when the alternative hypothesis is indeed true, or in other words, failing to reject the null when the null hypothesis is false. Type II error is commonly known as β. Beta relates to another important parameter in statistical testing which is *power*. Power is equal to (1-β) and is essentially the ability to avoid making a type II error. Like α, power is also defined by the researcher, and is typically set at 0.80. Below is a schematic of the relationships between α, β and power.

| Decision | Truth | |
|---|---|---|
| | Null True | Null False |
| Reject Null | α | power |
| Accept Null | | β |

### Students' T test

This test is most commonly used to test the difference between the means of the dependent variables of two groups. For example, this test would be appropriate if one wanted to evaluate whether or not a new anti-hypertensive drug reduces mean systolic blood pressure.

### Example

To evaluate if drug Z reduces mean systolic blood pressure, a randomized clinical trial will be performed where 12 individuals receive drug Z and 8 receive a placebo. The null hypothesis to be tested is that there is no difference in the mean systolic blood pressure of the experimental and placebo groups. The alternative hypothesis is that there is a difference between the means of the two groups. The type I error for your trial will be 5%.

### Results

Below is the group assignments and resulting systolic blood pressure (SBP)

| Patient | Assignment | Systolic BP |
|---|---|---|
| 1 | Drug Z | 100 |
| 3 | Drug Z | 110 |
| 5 | Drug Z | 122 |
| 7 | Drug Z | 109 |
| 9 | Drug Z | 108 |
| 11 | Drug Z | 111 |
| 13 | Drug Z | 118 |
| 15 | Drug Z | 105 |
| 17 | Drug Z | 115 |
| 18 | Drug Z | 119 |
| 19 | Drug Z | 106 |
| 20 | Drug Z | 109 |
| 2 | Placebo | 129 |
| 4 | Placebo | 125 |
| 6 | Placebo | 136 |
| 8 | Placebo | 129 |
| 10 | Placebo | 135 |
| 12 | Placebo | 134 |
| 14 | Placebo | 140 |
| 16 | Placebo | 128 |

$$\text{mean}_{drug} = \frac{100 + 110 + \ldots + 109}{12} = 111 \text{ mm Hg}$$

$$\text{mean}_{placebo} = \frac{129 + 125 + \ldots + 128}{8} = 132 \text{ mm Hg}$$

$$\text{mean}_{drug} - \text{mean}_{placebo} = -21 \text{ mm Hg}$$

Now that we have determined the difference between means, we need to determine the standard error for that difference which is calculated using the pooled estimate of the variance ($\sigma^2$).

The formula for the standard error of the drug Z group is:

$$\sigma^2 drug = \frac{\sum (SBP_{drug} - \text{mean}_{drug})^2}{n_{drug} - 1} =$$

$$\sigma^2 drug = \frac{[(100-111)^2 + (110-111)^2 + \ldots + (109-111)^2]}{12-1} = 40.9$$

The standard error for the placebo group is calculated in the same manner substituting the values for the placebo group.

$$\sigma^2 placebo = 25.1$$

Next, we would need to calculate a pooled estimate of the variance using the following equation:

$$\sigma^2_p = \frac{[(n_{drug} - 1)\ \sigma^2_{drug}] + (n_{placebo} - 1)\ \sigma^2_{placebo}]}{(n_{drug} - 1) + (n_{placebo} - 1)} =$$

$$\sigma^2_p = \frac{(11)(40.9) + (7)(25.1)}{11 + 7} = \frac{626}{18} = 34.8$$

The pooled estimate of the variance can then be utilized to calculate the standard error for the difference in means:

$$SE^2 (\text{mean}_{drug} - \text{mean}_{placebo}) = \frac{\sigma^2_p}{n_{drug}} + \frac{\sigma^2_p}{n_{placebo}}$$

$$SE^2 = \frac{34.8}{12} + \frac{34.8}{8} = 7.236$$

$$SE = 2.69$$

Now we are finally ready to test for significant differences in the mean blood pressure of our two groups: (*mean indicates the hypothesized values for the null-generally this quantity would = 0 when there is no difference expected between the drug and placebo groups).

$$t = \frac{(\text{mean}_{drug} - \text{mean}_{placebo}) - (*\text{mean}_{drug} - *\text{mean}_{placebo})}{SE\ (\text{mean}_{drug} - \text{mean}_{placebo})}$$

$$t = \frac{-21 - 0}{2.69} = -7.8 = |-7.8| = 7.8$$

We now compare our calculated value to a table of critical values for the Students' T distribution (found in most basic statistics books). The table also requires that we know the degrees of freedom and the value of a we have selected. Degrees of freedom (df) refers to the amount of information that a sample has in estimating the variance. It is generally the sample size minus one. The df for our calculation is 12 + 8 - 2 = 18 (the sample size for each group - 1). With a two tailed a of 0.05, our value |-7.8| is greater than the critical value from the table (2.101). Thus, we can reject the null hypothesis that there is no difference between mean blood pressure levels, and accept, by elimination, our alternative hypothesis.

### Chi-square analysis

What happens if we don't have continuous data, and are faced with categorical data instead? We could turn to chi-square analysis to evaluate if there are significant associations between a given exposure and outcome (the row and column variables in a contingency table). 2 X 2 contingency tables are one of the most common ways to present categorical data, and we can see this in analyzing data that was collected to address the question presented in this notebook.

Is a visit with a social worker, in addition to regular medical visits, associated with greater satisfaction of care for cancer patients as compared to those who only have regular medical visits?

Below is a generic 2 X 2 table representing the data. It is important to note the set-up of the table, as cell "a" generally represents the group of interest (diseased and exposed) and cell d represents the referent group (no disease and unexposed).

| | Row value (often disease or health outcome) | | |
|---|---|---|---|
| **Column Value (often Exposure)** | 1 | 0 | **Total** |
| 1 | *a* | *b* | *a + b* |
| 0 | *c* | *d* | *c + d* |
| **Total** | *a + c* | *b + d* | *n* |

Here we have the contingency table with data from our trial:

| | Greater Satisfaction? | | |
|---|---|---|---|
| **Social Worker Visit?** | Yes | No | **Total** |
| Yes | 64 | 46 | 110 |
| No | 36 | 54 | 90 |
| **Total** | 100 | 100 | 200 |

In chi-square analysis we are testing the null hypothesis that there is no association between a social worker visit and a greater satisfaction with care.

Generally, in evaluating this type of data, it is important for each of the individual cells to have large values, (i.e. greater than 5 or 10 each), If these conditions are not met, a special type of chi-square analysis is conducted called the Fisher's exact test. This will not be discussed in this notebook.

To calculate the chi-square statistic ($\chi^2$ ):

$$\chi^2 = \sum \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

with *i* representing the frequency in a particular cell of the 2 X 2 table. Below is the calculation for the frequencies that are **expected** in each cell.

| | Row value | | |
|---|---|---|---|
| **Column Value** | 1 | 2 | **Total** |
| 1 | $\frac{(a+b)(a+c)}{n}$ | $\frac{(a+b)(b+d)}{n}$ | $a + b$ |
| 2 | $\frac{(c+d)(a+c)}{n}$ | $\frac{(c+d)(b+c)}{n}$ | $c + d$ |
| **Total** | $a + c$ | $b + d$ | $n$ |

Thus, we now have a table that has both the actual and expected (in parentheses) values:

| | Greater Satisfaction? | | |
|---|---|---|---|
| **Social Worker Visit?** | Yes | No | **Total** |
| Yes | 64 (55) | 46 (55) | 110 |
| No | 36 (45) | 54 (45) | 90 |
| **Total** | 100 | 100 | 200 |

With this information, we can now calculate the $\chi^2$ statistic:

$$\chi^2 = \sum \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

$$\chi^2 = \frac{(64-55)^2}{55} + \frac{(46-55)^2}{55} + \frac{(36-45)^2}{45} + \frac{(54-45)^2}{45}$$

$$\chi^2 = 6.545$$

The chi-square statistic for these data has approximately 1 degree of freedom, an α of 0.05, and it is compared to the critical values on standard Chi-square table. Note that the degrees of freedom would increase as the number of rows and columns of our tables increases (for instance a 3 X 4 table). Since our calculated value ($\chi 2 = 6.545$) is greater than the critical value (3.841), we can once again reject the null hypothesis that there is no association between the exposure and the outcome of interest, and conclude that in this case seeing a social worker is significantly associated with a greater satisfaction with care.

**Important notes**

It is important to remember that the statistical tests and examples presented here are only an elementary presentation of the large scope of situations that can be addressed by these data. The intention of this notebook is to provide a basic understanding of the underlying principles of these statistical tests rather than implying that what has been presented is appropriate for every situation.

Further information about these statistical tests and other applications can be found in the following references:

Statistical First Aid: Interpretation of Health Research Data by Robert P Hirsch and Richard K. Riegelman. Blackwell Scientific Publications, Cambridge, MA 1992.

Categorical Data Analysis, Using the SAS System by ME Stokes, CS Davis, and GG Koch. SAS Institute Inc., Cary, NC, 2001.

## Practice Questions

*Answers are at the end of this notebook*

Researchers are conducting a study of the association between working in a noisy job environment and hearing loss. The researchers' null hypothesis is that there is no difference in hearing loss between people who work in a noisy job environment compared with people who work in a quiet job environment. The researchers' alternative hypothesis is that there is a difference in hearing loss between people who work in a noisy job environment compared with people who work in a quiet job environment. The researchers decided to set their alpha level at 0.05. The researchers' analysis results show a p-value of 0.0003 (please note that for the purposes of this question you are being provided with just the p-value from the study when in reality a study analysis is much more complex).

1) True or false: The alpha level of 0.05 is an arbitrary value.

2) True or false: Based on the results, the researchers can conclude their null hypothesis is true.

3) True or false: Based on these results, the researchers should reject the assumption that their null hypothesis is true.

4) True or false: An alpha level of 0.05 means there is a 0.05 percent chance that the researchers will incorrectly reject the null hypothesis.

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

**Answers to Practice Questions**

1) True or false: The alpha level of 0.05 is an arbitrary value.

Answer: True

This statement is true. The level of alpha is often set at 0.05, however, this choice is arbitrary and researchers may choose a different value.

2) True or false: Based on the results, the researchers can conclude their null hypothesis is true.

Answer: False

This statement is false. Researchers should never conclude that their null hypothesis is true. It is possible to conclude that we should fail to reject the assumption that the null hypothesis is true but this is not the same as concluding that the null hypothesis is actually true.

3) True or false: Based on these results, the researchers should reject the assumption that their null hypothesis is true.

Answer: True

This statement is true. The p-value was 0.0003 which is less than the alpha of 0.05. The researchers should reject their null hypothesis that there is no difference in hearing loss diagnosis between people who work in a noisy job environment compared with people who work in a quiet job environment.

4) True or false: An alpha level of 0.05 means there is a 0.05 percent chance that the researchers will incorrectly reject the null hypothesis.

Answer: False

This statement is false. An alpha level of 0.05 means there is a 5% chance the researchers will incorrectly reject the null hypothesis.

# ERIC Notebook

Second Edition

## Confounding Bias, Part I


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Confounding is one type of systematic error that can occur in epidemiologic studies. Other types of systematic error such as information bias or selection bias are discussed in other ERIC notebook issues.

Confounding is an important concept in epidemiology, because, if present, it can cause an over- or under-estimate of the observed association between exposure and health outcome. The distortion introduced by a confounding factor can be large, and it can even change the apparent direction of an effect. However, unlike selection and information bias, it can be adjusted for in the analysis.

### What is confounding?

Confounding is the distortion of the association between an exposure and health outcome by an extraneous, third variable called a confounder. Since the exposure of interest is rarely the only factor that differs between exposed and unexposed groups, and that also affects the health outcome or disease frequency, confounding is a common occurrence in etiologic studies.

Confounding is also a form a bias. Confounding is a bias because it can result in a distortion in the measure of association between an exposure and health outcome.

Confounding may be present in any study design (i.e., cohort, case-control, observational, ecological), primarily because it's not a result of the study design. However, of all study designs, ecological studies are the most susceptible to confounding, because it is more difficult to control for confounders at the aggregate level of data. In all other cases, as long as there are available data on potential confounders, they can be adjusted for during analysis.

Confounding should be of concern under the following conditions:

1. Evaluating an exposure-health outcome association.
2. Quantifying the degree of association between an exposure and health outcome. For example, you might want to quantify how being overweight increases the risk of cardiovascular disease (CVD). If you were concerned about age as a confounder, you would “control for” the effect of age in your statistical modeling.

In one study, the rate ratio might change from 4.0 to 3.7 when controlling for age, whereas in another study, a rate ratio of 4 may change to 1.2 after controlling for age.

3. Multiple causal pathways may lead to the health outcome. If there is only one way to contract the health outcome or disease, confounding cannot occur. This criterion is almost always met as health outcomes can inevitably be caused by different agents, different transmission routes, or different biological or social mechanisms.

A few examples of research questions in which you would want to consider confounding are listed below:

1. Does being overweight increase the risk of coronary heart disease (CHD) -- independently of cholesterol, hypertension, and diabetes?
2. Does tobacco advertising entice adolescents to experiment with tobacco independently of whether or not their parents smoke?

### Assessing confounding

Each potential confounder has to meet two criteria before they can be confounders: *Criterion 1* is that the potential confounder must be a known risk factor for the health outcome or disease.

Broadly speaking, a risk factor is any variable that is:

1. Already known to be "causally related" to the health outcome or disease (though not necessarily a direct cause) AND
2. Antecedent to the health outcome or disease on the basis of substantive knowledge or theory, and/or on previous research findings.

The confounding factor must be predictive of the health outcome or disease occurrence apart from its association with exposure; that is, among unexposed (reference) individuals, the potentially confounding factor should be related to the health outcome or disease.

With an epidemiological data set, one can calculate whether or not a potential confounder is a risk factor using the following mathematical formula:

**Criterion 1 for confounding: mathematical formula**

Criterion 1 for confounding is the following: among the unexposed, there should be an association between the confounder and the health outcome.

To convert this to a mathematical equation, the first thing to realize is that Criterion 1 involves calculating a measure of association ("there should be an association between the confounder and the health outcome"). Examples of measures of association are: risk ratios, rate ratios, odds ratios, and risk differences – the type of measure depends on the type of data available, and the scale on which the measure of association is assessed (additive or multiplicative scale). This measure of association will be calculated among the unexposed population only.

For a prospective cohort study where we want to measure the association on a multiplicative scale, we will calculate the following rate ratio (RR):

RR CD/E- = risk ratio confounder in unexposed

$$\frac{\text{Rate of new cases among population A}}{\text{Rate of new cases among population B}}$$

where the rate of new cases = the number of new cases divided by the total number of susceptible individuals. Population A is comprised of all individuals who have the confounder (C+) but who are unexposed (E-), and population B is comprised of all individuals who don't have the confounder (C-) or the exposure (E-).

For a case-control study using odds ratios (OR), the formula for Criterion 1 is:

ORCD/E- = odds ratio confounder in unexposed

$$\frac{\text{Odds that cases have confounder among population F}}{\text{Odds that controls have confounder among population F}}$$

where the odds that the cases have the confounder = the number of cases with the confounder (C+) divided by the number of cases without the confounder (C-) and where population F is comprised of all individuals who are not exposed (E-).

Now that the risk ratio, rate ratio, or odds ratio for the association between the confounder and health outcome among the unexposed has been calculated, how is it interpreted?

For the confounder to be a risk factor, the measure of association has to be greater than 1 (for a harmful association), or less than 1 (for a protective association).

Age and smoking status, for example, are widely considered to be risk factors for lung cancer, even though the mechanisms by which both variables are determinants of this disease are not well understood. On the other hand, race is not considered to be a risk factor for lung cancer. Unnecessary adjustment of variables that are not confounders can lower precision and may even introduce bias into the estimate of effect.

*Criterion 2* is that the potential confounder must be associated with the main exposure, but not as a result of the exposure. In other words, all potential confounders should be working independently and not as part of the proposed exposure-health outcome pathway. One can calculate whether or not a potential confounder is associated with the main exposure using a mathematical formula.

**Criterion 2 for confounding: mathematical formula**

Criterion 2 for confounding is the following: the distribution of the confounding variable differs between exposed and unexposed groups.

To convert this to a mathematical equation, the first thing to realize is that Criterion 2 involves calculating a measure of association.

For a prospective cohort study, we will calculate the following risk ratio:

$RR_{EC}$=

$$\frac{\text{\% individuals with confounder (C+) among Population A}}{\text{\% individuals with confounder (C+) among Population B}}$$

where Population A will be comprised of all individuals who are exposed (E+), and where Population B will be comprised of all individuals who are unexposed (E-).

For a case-control study using odds ratios (OR) the formula for Criterion 2 is:

$OR_{EC}$=

$$\frac{\text{Odds of controls having the confounder (C+) among Population A}}{\text{Odds of controls having the confounder (C+) among Population B}}$$

where odds of controls having the confounder (C+) = number of controls having the confounder (C+) divided by the number of controls not having the confounder (C-). Population A is comprised of all individuals who are exposed (E+), and population B is comprised of all individuals who

are unexposed (E-). Note the additional inclusion criteria for case-control studies: the individuals included in this calculation must include only those who have the potential to be cases (the control group).

Now that the risk ratio, rate ratio, or odds ratio for the association between the confounder and exposure has been calculated, how is it interpreted? For the confounder associated with the exposure, this association has to be greater than 1 (for a harmful association) or less than 1 (for a protective association).

To decide whether a variable is working independently of the association of interest, there must be a biological or social mechanism to causally link the exposure of interest to the disease or health outcome. Such decisions should be made on the basis of the best available information, including non-epidemiological (i.e., clinical, sociological, psychological, or basic science) data. This criterion is obviously satisfied if the confounding factor precedes the exposure and health outcome or disease.

For instance, if interested in assessing the association between physical inactivity and cardiovascular disease (CVD), body weight should not be controlled for if being overweight may be an intermediary step in the causal pathway between physical inactivity and CVD.

In contrast, if the proposed causal pathway is independent of body weight, then body weight can be considered a potential confounder. If intervening variables are controlled for in the analysis, it may reduce or eliminate any indications in the data of a true association between disease and exposure.

ERIC Notebook *Confounding Bias Part II and Effect Measure Modification,* discuss control of confounders in epidemiological studies.

**Terminology**

*Confounding bias:* A systematic distortion in the measure of association between exposure and the health outcome caused by mixing the effect of the exposure of primary interest with extraneous risk factors.

**Practice Questions**

*Answers are at the end of this notebook*

Researchers have conducted a cohort study in country A to examine the association between a diet high in fat and the risk of colon cancer. The researchers believe that vitamin use may be a confounder. Use the 2x2 tables below to determine if vitamin use is a confounder in the high fat diet-colon cancer association.

| | Colon cancer | No colon cancer | Total |
|---|---|---|---|
| Exposed to a high fat diet | 254 | 2220 | 2474 |
| Not exposed to a high fat diet | 150 | 1500 | 1650 |

Among people exposed to a high fat diet (n=2474):

| | Colon cancer | No colon cancer | Total |
|---|---|---|---|
| Takes daily vitamin | 150 | 1830 | 1980 |
| Does not take daily vitamin | 104 | 390 | 494 |

Among people <u>not</u> exposed to a high fat diet (n=1650):

| | Colon cancer | No colon cancer | Total |
|---|---|---|---|
| Takes daily vitamin | 50 | 800 | 850 |
| Does not take daily vitamin | 100 | 700 | 800 |

1) Is vitamin use an independent risk factor or protective factor for colon cancer?

2) Is vitamin use differentially distributed between the high fat diet and low fat diet groups?

3) Compare the crude risk ratio with the risk ratios stratified by vitamin use.

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the

## Answers to Practice Questions

1) Risk ratio of vitamin users getting colon cancer among the non-exposed group: (50/850) / (100/800)= 0.47

A risk ratio of 0.47 shows that vitamin use is a moderate inverse predictor of colon cancer. In this study population, vitamin use was protective for colon cancer.

2 Among people who eat a high fat diet there are 1980/2474= 80% vitamin users

Among people who do not eat a high fat diet there are 850/1650= 52% vitamin users

So vitamin use is differentially distributed among the high fat and low fat diet exposure groups.

3) The crude risk ratio (not stratified by vitamin use) is the risk of colon cancer from high fat diet exposure / the risk of colon cancer from low fat diet exposure. Crude risk ratio = (254/2474) / (150/1650) = 1.13

The risk ratio for colon cancer among vitamin users with a high fat diet is:
Risk ratio = (150/1980) / (50/850)= 1.29

The risk ratio for colon cancer among non-vitamin users with a high fat diet is:

Risk ratio = (104/494) / (100/800) =1.68

The crude risk ratio of 1.13 and the vitamin-specific risk ratio of 0.47 (from question 1) are not in between the stratified risk ratios, they are both lower than the stratified risk ratios. Thus, the crude risk ratio is confounded by vitamin use.

# ERIC Notebook

Second Edition

## Confounding Bias, Part II and Effect Measure Modification


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

**Confounding is one type of systematic error that can occur in epidemiologic studies, and is a distortion of the association between an exposure and health outcome by an extraneous, third variable**

**Effect measure modification (EMM) is when a measure of association, such as a risk ratio, changes over values of some other variable.** In contrast to confounding which is a distortion, EMM is of scientific interest ,answers a research question, and can help identify susceptible or vulnerable populations. Are children with poor nutrition, exposed to water pollution, at higher risk of gastrointestinal symptoms compared with children with good nutrition who are exposed to water pollution?

Both confounding and EMM can be considered in the design phase of research or in the analysis phase. We will discuss both in this notebook.

### Calculating and adjusting for confounding

The previous issue of ERIC Notebook, "Confounding Bias, Part I", discussed two criteria for identifying potential confounders in a study. Once potential confounders have been identified, the next step is to evaluate if and how much the confounders bias the study results. To do this, results where confounding is ignored, the *"crude"* measure of association, are compared to results that have been corrected for distortions due to confounding, the *"adjusted"* measure of association.

This methodology makes 2 assumptions:

First, the data are obtained by simple random sampling rather than by some more restrictive subject selection procedure, like matching.

The second assumption is that the exposure, health outcome, and confounder variables are all dichotomous (i.e., having only two strata). If the variables are in a continuous format, they can either be dichotomized, or they must be adjusted for simultaneously to calculate the true measure of association.

Methods to calculate adjusted measures of associations differ by the need to control each confounder individually or all confounders simultaneously.

Before calculating an adjusted measure of association using stratified analyses, one must first assess the presence of *effect measure modification* (EMM). When effect measure modification is present, it can be difficult to ascertain whether or not confounding is occurring.

Two examples of EMM are:

- A breast cancer education program (the exposure) that is much more effective in reducing breast cancer in rural areas than urban areas. Here, the area (rural or urban) is an effect measure modifier.
- The finding that a reduction in regional public transportation services (the exposure) affects individuals with little or no access to a car much more than those individuals with access to a car. In this example, having access to a car is the effect measure modifier.

**What is effect measure modification?**

When estimates of an exposure-health outcome relationship stratified by a confounder are sufficiently different from one another (i.e., risk ratio level 1=4.0 and risk ratio level 2 =0.2), they suggest that two different exposure-health outcome relationships may be operating, one in each level of the confounder.

EMM is different from confounding, where instead of "competing" with the exposure of interest in explaining the etiology of a health outcome or disease, the effect measure modifier identifies subpopulations that are particularly susceptible to the exposure of interest.

**Is effect measure modification present?**

To calculate whether EMM may be occurring in the study, first calculate three measures of association:

A = The overall, crude measure of association of the exposure-health outcome

B1 = The measure of the exposure-health outcome association among all study participants who have a history of the confounding variable (C+)

B2 = The measure of the exposure-health outcome association among all study participants who do not have a history of the confounding variable (C-)

Use the Figure 1 below as a guide on how to interpret the meaning of these three measures of associations.

As a general rule, if B1 and B2 are basically equal in value, but different from A, then confounding is present and EMM is not present. EMM is present when B1 and B2 are different from one another, and at least one (B1 or B2) is different from A. Both EMM and confounding can occur simultaneously.

If stratified analysis is used to adjust for EMM, confounders should be addressed using more complex statistical techniques, as stratifying results on more than

Figure 1

Risk Ratio Example

one variable splits the sample into significantly smaller sample sizes and limits generalizability.

**Calculating adjusted summary estimates**

If no EMM is present, then stratum-specific estimated effects can be pooled to form a summary estimate of effect across strata. This summary estimate represents an adjusted risk ratio (a risk ratio adjusted for confounding).

Although there are many ways to calculate the adjusted risk ratio, the Mantel Haenszel procedure, is the most common pooling procedure.

**Calculating adjusted measures when all confounders are assessed simultaneously**

Simultaneous control of two or more variables can give different (and potentially more interesting) results from those obtained by controlling for each variable separately. Simultaneous control of confounders better emulates the natural environment where exposures, diseases, and confounders of interest are found, than does individual control of confounders.

Simultaneous control of several confounders to calculate adjusted measures is done through mathematical modeling.

To control for confounding using mathematical modeling, simply include the confounding variables as independent variables in the model. The simplicity of this method of adjustment for confounding is one of the attractive features of using mathematical models in epidemiology.

Although many types of mathematical models are available, there is generally only one type of model that is appropriate for the goals of a specific data analysis and for the type of data available. The most common mathematical model used in epidemiology is logistic regression.

This model has the general format of

$$y = a + b_{1x} + b_{2z2} + \ldots + b_{izi}$$

Individual-level data need to be provided on:

1. y: the health outcome in a dichotomous format
2. x: the exposure
3. z2 to zi: the confounders

With this information and statistical analysis software, researchers can then calculate the following:

1. a: the y-axis intercept
2. b1: the coefficient for the exposure variable
3. b2 to bi: the coefficients for each confounder that is controlled for in the model.

These coefficients (except for a) are very useful, as they can be transformed into odds ratios. The odds ratio obtained from b1 of this model is an interpretable measure of association describing the relationship between the exposure (x) and the health outcome (y) after adjustment for confounding variables ($z_2$ to $z_i$).

The biggest disadvantages to using mathematical models are the assumptions that must be met by the dataset in order to use them – often the data may not conform to all assumptions. It is advisable to do regression diagnostics sometime during the data analysis stage to check these assumptions.

**Is adjustment for confounding necessary?**

If the adjusted effects are markedly different from the crude effect (typically a 10 or 15% change from crude to adjusted), then confounding is present and should be controlled for. Researchers should report the cut-off they used for their analysis (e.g. 5%, 10%, or 50%) in the selection of confounders for adjustment.

If the adjustment of confounding variables changes the results only slightly (less than 10%), then the tendency would be to ignore this influence, since the more variables controlled for, the less precise (the wider the confidence intervals) the study results will be. The benefits of ignoring the minor confounders would outweigh the costs.

Also consider whether it is important to control for potential confounders such as age, simply because many

readers would not trust results that are not adjusted for age. This distrust stems from knowledge that age is strongly related to disease and mortality rates (similar comments would apply to sex).

**Control of confounding**

*In the analysis phase:*

Once data have been collected, there are two options for control of confounding: Stratified analysis or mathematical modeling. Both methods were described above when calculating the effect of confounding on the measure of association. Briefly, stratified analysis pools the measure of association calculated in each strata of the confounder into one summary estimate. Mathematical modeling uses a more complex approach and makes more assumptions than stratified analysis.

*In the design phase:*

Some confounders should be controlled for in the study design stage of a study, rather than in the analysis stage. It may be necessary to do this if the confounder is very strong and when the anticipated sample size will be large enough to deal with it in the design stage. Some study designs are more favorable for controlling for confounding than others.

Restriction, matching, and randomization are common techniques used to minimize confounding in the design phase. These techniques are not exclusive to one another. Several different control methods may be used at once.

*Restriction*

Confounding can be controlled for by restricting the study population to those who are unexposed to one or more confounding variables. An example of restriction is to restrict a study population to nonsmokers when studying the association of environmental radon with lung cancer. Restriction is ideal when the exposure-health outcome relationship has strong confounders because it can be an efficient, convenient, inexpensive, and straight-forward method of controlling for confounding. However, the restricted variable, for instance smoking in the given example, cannot be assessed for confounding. Restriction may not always be logistically feasible because the sample size of available study participants is decreased, sometimes to the point that a study cannot be done.

**Example**

For instance, a group of 30 HIV-positive skydivers of varying age (20% twenty-something, 30% thirty-something, 50% senior citizens) has been identified with which to study behavioral risk factors for HIV infection, independent of age. Therefore, the control group should be 30 HIV-negative skydivers with the same age distribution as the group of HIV-positive skydivers (20% twenty-something, 30% thirty-something, 50% senior citizens). This would be accomplished through matching the controls to the cases by age, by selecting only HIV-negative skydivers who contribute to the pre-determined age distribution.

*Matching*

Confounding can also be controlled through matching on the confounder variable(s). Matching involves constraining the control group (for case-control studies) or the unexposed group (for cohort studies) such that the distribution of the confounding variable(s) within these groups are similar (or identical) to the corresponding distribution within the index group (the case group for case -control studies or the exposed group for cohort studies). Matching can be viewed as imposing a "partial restriction" on the values of the confounding variables, since only the control or unexposed group is restricted.

Analysis of matched data requires special consideration, because the control, or unexposed, group is not a random sample of study participants; they should be considered to be a biased sample. Techniques for analyzing matched data include conducting the data analysis separately for each level of the confounder (stratified analysis) and using conditional logistic regression.

When considering matching, consider four factors:

1. Precision (generally increased with matching)
2. Cost (generally lowered with matching, because a smaller sample size is needed)

3. Feasibility (can be increased with matching)
4. Flexibility in deciding whether to match

Also keep in mind that variables matched cannot be assessed for confounding.

*Randomization*

Randomization is an ideal method for controlling for confounding because this method can control both known and unknown confounders. However, because randomization requires that the exposure status of individuals be assigned to study participants, observational study designs such as cross-sectional, cohort, case-control and ecological studies cannot use randomization to control for confounding. For controlled clinical trials however, randomization is a common method to control for confounding.

Use of randomization to control for confounding presumes that random classification of individuals into groups will produce groups that have an equal (or similar) distribution of confounders. For example, the theory of randomization says that given two randomly selected groups of students, each group will have an equal percentage of females, an equal percentage of individuals with white-colored tops, an equal percentage of brown-eyed individuals, and so forth. Thus, randomization, if done correctly, will produce homogeneous groups of individuals. When an exposure is applied to one of these homogeneous groups, but not to the other, the only difference between the two groups is their exposure status. In this situation, confounding, the unequal distribution of a risk factor between exposed and non-exposed groups, cannot occur.

**Terminology**

*Effect measure modification:* a variation in the magnitude of a measure of exposure effect across levels of another variable

*Randomization:* random assignment of subjects to exposure categories

*Matching:* the selection of controls, or unexposed subjects, that are identical, or nearly so, to the cases, or exposed subjects, with respect to the distribution of one or more potentially confounding factors

From: Modern Epidemiology, Rothman KJ and Greenland S, 1998

The key to proper control of confounding through randomization is having a sufficiently large sample size in each randomized group. Rothman and Greenland (1998) state that having at least 50 subjects, preferably 100 or more, will usually assure that potential confounders are equally distributed among each study group.

**Practice Questions**
*Answers are at the end of this notebook*

1) There is a vaccination campaign in effect in some counties of a state to educate parents about childhood diseases and recommended vaccines for their children. Researchers find that the vaccination campaign seems to be effective. Children who live in the counties with the vaccination campaign are 4 times as likely to be vaccinated according to recommendations when compared with children who live in counties without the vaccination campaign. However, even looking only within counties with the vaccination campaign, researchers find that children of families with incomes above the poverty level are twice as likely to vaccinate their children when compared with children of families with incomes below the poverty level. Which of the following can be determined from this example?

Choose all that apply:

a) Family income appears to be a confounder
b) Family income appears to be an effect measure modifier
c) The vaccination campaign appears to be effective
d) The vaccination campaign is the study outcome
e) The vaccination campaign is the study exposure

2) Researchers conducted a cohort study in country B to examine the association between a diet high in fat and the risk of colon cancer. The researchers believe that vitamin use may be a confounder but they first need to examine whether or not vitamin use is an effect measure modifier. Use the 2x2 tables below to determine if vitamin use is an

effect measure modifier or a confounder in the high fat diet- colon cancer association.

| | Colon cancer | No colon cancer | Total |
|---|---|---|---|
| Exposed to a high fat diet | 254 | 2220 | 2474 |
| Not exposed to a high fat diet | 98 | 2300 | 2398 |

Among people exposed to a high fat diet:

| | Colon cancer | No colon can- | Total |
|---|---|---|---|
| Takes daily | 150 | 1830 | 1980 |
| Does not take | 104 | 390 | 494 |

Among people not exposed to a high fat diet:

| | Colon cancer | No colon cancer | Total |
|---|---|---|---|
| Takes daily vitamin | 80 | 1500 | 1580 |
| Does not take daily vitamin | 18 | 800 | 818 |

a) Is vitamin use an independent risk factor for colon cancer?

b) Is vitamin use differentially distributed between the high fat diet and low fat diet groups?

c) Compare the crude risk ratio with the risk ratios stratified by vitamin use.

## References

Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. Belmont, CA: Lifetime Learning Publications, 1982.

Miettinen OS, Cook EF. Confounding: essence and detection. American Journal of Epidemiology 114(4):593-603, 1981 Oct.

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

Rothman KJ. Introduction to Epidemiology. New York, NY: Oxford University Press, 2002

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. Answer choices b, c and e are correct. In this example, family income appears to be an effect measure modifier because the effect of the vaccination campaign differs markedly between the 2 sub-populations (family income above versus below poverty level). The vaccination campaign does appear to be effective since children in counties with the vaccination campaign were 4 times as likely to be vaccinated as children in counties without the vaccination campaign. In this example, the researchers are studying the outcome of whether or not children are vaccinated according to recommendations. The vaccination campaign is the exposure.

2.

a) Is vitamin use an independent risk factor for colon cancer?

Answer:
Risk ratio of vitamin users getting colon cancer among the non-exposed group:
(80/1580) / (18/818)=2.3

Yes, a risk ratio of 2.3 shows that vitamin use is a moderate predictor of colon cancer.

b) Is vitamin use differentially distributed between the high fat diet and low fat diet groups?

Answer:
Among people who eat a high fat diet there are
1980/2474= 80% vitamin users
Among people who do not eat a high fat diet there are
1580/2398= 66 % vitamin users
So vitamin use is differentially distributed among the high fat and low fat diet exposure groups.

c) Compare the crude risk ratio with the risk ratios stratified by vitamin use.

Answer:
The crude risk ratio (not stratified by vitamin use) is the risk of colon cancer from high fat diet exposure / the risk of colon cancer from low fat diet exposure.
Crude risk ratio = (254/2474) / (98/2398) = 2.51

The risk ratio for colon cancer among vitamin users with a high fat diet is:
Risk ratio = (150/1980) / (80/1580)= 1.50

The risk ratio for colon cancer among non-vitamin users with a high fat diet is:
Risk ratio = (104/494) / (18/818) = 9.6

The crude risk ratio of 2.51 and the vitamin-specific risk ratio of 2.3 (from question 1) are both in between the stratified risk ratios. The 2 stratified risk ratios differ greatly from one another (9.6 for non-vitamin users versus 1.5 for vitamin users). This example shows the presence of effect measure modification.

# ERIC Notebook

Second Edition

## Cross-sectional Studies


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Like cohort studies, cross-sectional studies conceptually begin with a population base. But unlike cohort studies, in cross-sectional studies we do not follow individuals over time. Instead, we only look at the prevalence of disease and/or exposure at one moment in time. These studies take a "snapshot" of the proportion of individuals in the population that are, for example, diseased and non-diseased at one point in time. Other health outcomes besides diseases may also be studied. Cross-sectional studies also differ from cohort studies in the populations that are studied. Cohort studies begin by selecting a population of persons who are at risk of for a specific disease or health outcome; cross-sectional studies begin by selecting a sample population and then obtaining data to classify all individuals in the sample as either having or not having the health outcome.

| | Cohort | Cross-sectional |
|---|---|---|
| Study group | Population-at-risk | Entire population (or a sample) |
| Common Measures | Risks and Rates | Prevalence |

### Ways to use cross-sectional studies

Cross-sectional studies are used both descriptively and analytically. *Descriptive cross-sectional studies* simply characterize the prevalence of a health outcome in a specified population. Prevalence can be assessed at either one point in time (*point prevalence*) or over a defined period of time ( *period prevalence*). Period prevalence is required when it takes time to accumulate sufficient information on a disease in a population, e.g. what proportion of persons served by a public health clinic over a year have hypertension. These prevalence measures are commonly used in public health; often the point or period aspect is not specified.

In *analytical cross-sectional studies,* data on the prevalence of both exposure and a health outcome are obtained for the purpose of comparing health outcome differences between exposed and unexposed.

Analytical studies attempt to describe the prevalence of, for example, disease or non-disease by first beginning with a population base. These studies differ from solely descriptive cross-sectional studies in that they compare the proportion of

exposed persons who are diseased (a/(a+b)) with the proportion of non-exposed persons who are diseased (c/(c+d)).

### Calculating prevalence

The prevalence of a health outcome is simply the proportion of individuals with the health outcome in a population.

**Prevalence = cases / total population**

For the following example, two different sub-measures of prevalence can be calculated: the prevalence of coronary heart disease (CHD) among the exposed (people who are not active) and the prevalence of CHD among the unexposed.

**Example:**

| | Present | | Absent CHD | | Total |
|---|---|---|---|---|---|
| Not active | 50 | a | b | 200 | 250 |
| Active | 50 | c | d | 700 | 750 |
| Total | 100 | | | 900 | 1000 |

$P_1$= a/a+b= 50/250 = 20.0% prevalence of CHD among people who are not active.

$P_0$= c/c+d = 50/750 = 6.7% prevalence of CHD among people who are active.

### The prevalence odds ratio

The prevalence odds ratio (POR) is calculated in the same manner as the odds ratio.

**POR = ad / bc**

### The prevalence ratio

The prevalence ratio (PR) is analogous to the risk ratio (RR) of cohort studies. The denominators for both ratios are fixed populations – fixed at the start of the study in the case of a cohort study, and fixed at the point or period of time for the case-control study. The PR is similar to a RR when the outcome occurs over a short period of time. For example, one would calculate a prevalence ratio for an acute outbreak of tuberculosis in a prison population. This is in contrast to calculating the overall prevalence of positive tuberculin skin tests among the prisoners.

The prevalence ratio can also be calculated from the information on CHD and physical activity:

PR = (a/N1) / (c/N0)

PR= (50/250) / (50/750) = 3.0

In this case, a prevalence ratio of 3.0 can be interpreted to mean that the people who are not physically active are 3 times as likely as those who are physically active to have CHD.

Note that it is preferable to calculate the prevalence odds ratio when the period for being at risk of developing the outcome extends over a considerable time (months to years) as it does in this example. **POR vs. PR**

For chronic disease studies or studies of long-lasting risk factors, POR is the preferred measure of association in cross-sectional studies. For acute disease studies, PR is the preferred measure of association. If the prevalence of disease is low, i.e. 10% or less in exposed and unexposed populations, POR = PR. Since cross-sectional studies are particularly useful for investigating chronic diseases (e.g. prevalence of AIDS) where the onset of disease is difficult to determine, or for studying long lasting risk factors (such as smoking, hypertension, and high fat diets), the prevalence odds ratio will generally be the preferred measure of association.

### Limitations of cross-sectional studies to evaluate risk

Recall that, under steady conditions, the prevalence of disease is influenced both by incidence and duration of disease (or survival with disease).

**Prevalence = Rate x Average Duration of Disease**

Persons who survive longer with a disease will have a higher probability of being counted in the numerator of a prevalence proportion. Short-term survivors will be less likely to be counted as a case. Incidence is influenced only by exposure, whereas prevalence is influenced both by exposure and duration of disease.

If exposure influences survival time, then the POR or PR will not provide a valid estimate of the risk ratio or rate ratio. Thus, the interpretation of the POR or PR is subject to survival bias.

Even if incidence remains constant, either an improvement in disease treatment (that results in higher cure rates) or increased lethality (resulting in a higher case fatality rate) will result in decreased prevalence. The disease itself or the threat of developing the disease may cause outmigration of cases from an environment perceived as causing disease, e.g. workers affected by toxic exposures in a plant may quit, while more resistant workers will stay. This selective migration can bias measures of prevalence.

### Other problems with interpretation of cross-sectional studies

Cross-sectional studies as well as case-control studies are affected by the *antecedent-consequent bias*, similar to the chicken and egg question (i.e. "which came first?"). This bias occurs when it cannot be determined that exposure preceded disease, since both are ascertained at the same time (unlike cohort studies or clinical trials). Antecedent-consequent bias does not affect cohort studies because subjects in cohort studies are selected for study because they are disease-free. Exposure is actually observed to precede disease only in a cohort design, including randomized trials.

### Uses of cross-sectional studies

Descriptive studies are an important method to evaluate the proportion of a population with disease or with risk factors for disease, such as the prevalence of asthma in children or the prevalence of elevated blood lead in toddlers.

Descriptive cross-sectional studies are widely used to estimate the occurrence of risk factors in segments of the population characterized by age, sex, race or socioeconomic status (SES). National examples of cross-sectional studies of great importance are the *decennial census* and the National Health and Nutrition Surveys (NHANES). Opinion polls and political polls are basically cross-sectional studies. Surveillance of changes in smoking habits or of other behavioral risk factors are sequential cross-sectional studies. The US National Health and Nutrition Examination Survey (NHANES) is one such example Similarly, surveillance of long lasting diseases such as AIDS is cross-sectional. Descriptive cross-sectional studies are useful for planning or administering preventive or health care services, surveillance programs, and surveys and polls.

Descriptive/analytical cross-sectional studies are useful for establishing preliminary evidence for a causal relationship. These studies are also useful for examining the association between exposure and disease onset for chronic diseases where researchers lack information on time of onset. Examples might include diet and arthritis, smoking and chronic bronchitis, and asthma and exposure to air pollution. Interpretation requires caution regarding potential association of duration of disease with exposure status *(survival bias)*.

Survival bias may be minimized if information can be obtained on exposures that clearly preceded the first symptoms of a chronic disease such as arthritis, diabetes, or chronic bronchitis. This depends on access to medical records before the onset of a chronic disease. In addition, it may be necessary to have historical records on an individual's exposure status prior to these first medical visits, e.g. where the person lived or where the person was employed.

### Terminology

*Antecedent-consequent bias*: occurs in cross-sectional studies when it cannot be determined if exposure preceded disease.

*Prevalence:* the proportion of diseased individuals in a population.

*Survival bias:* occurs in cross-sectional studies when the exposure influences survival time, and the distribution of that exposure will be distorted among a sample of survivors. (a.k.a. Neyman bias, incidence-prevalence bias, or selective survival bias)

### Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

### Practice Questions

*Answers are located at the end of this notebook*

1) Which of the following health outcomes could be studied using a cross-sectional study design? Choose all that apply.

a) The prevalence of diabetes among adults in the United States in 2014
b) The prevalence of diabetes among all patients seen at a particular health clinic on one day in 2014
c) The number of new cases of diabetes diagnosed among at risk adults in the United States in 2014
d) The number of people in a population with diabetes who are obese and the number of people in a population with diabetes who are not obese, in the United States in 2014

2) Researchers are studying HIV prevalence using a cross-sectional study design. Which of the following factors may affect the researchers' assessment of HIV prevalence in their study population? Choose all that apply.

a) Changes in HIV treatment
b) Changes in HIV virulence (HIV virulence refers to the ability of the virus to cause disease)
c) Population changes (e.g. migration) due to HIV infection, for example people who leave their community to live closer to physicians specialized in treating HIV
d) Other factors which may affect survival of HIV-infected persons (e.g. changes in healthcare practices)
e) Changes in HIV diagnostics
f) Changes in HIV awareness and education

### Resources:

Delgado-Rodriguez M and Llorca J. J Epidemiol Community Health. 2004;58:635–641.

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

### Answers to Practice Questions

1. Answer choices a, b and d are correct. A cross-sectional study design could be used to assess the prevalence of diabetes among adults in the United States in 2014. A cross-sectional study design could also be used to assess the prevalence of diabetes among all patients seen at a particular health clinic on one day in 2014 by simply recording which patients are known to have been diagnosed with diabetes divided by the number of total patients seen that day. However, the number of new cases of diabetes diagnosed among at risk adults in the United

States in 2014 could not be assessed using a cross-sectional study. The incidence of new cases cannot be measured using a cross-sectional "snapshot" in time design. Instead, a cohort study could be used to follow at risk participants over time to see which people were newly diagnosed with diabetes in 2014. The number of people in a population with diabetes who are obese and the number of people in a population with diabetes who are not obese could be assessed using a cross-sectional design, this would be an example of an analytical cross-sectional study.

2. Answer choices b, c, d, e and f are correct. Changes in HIV virulence could affect HIV-related mortality (survival time) and, thus, would affect prevalence. Any other factors which affect survival of HIV-infected persons would also affect prevalence. Migration could also affect assessment of HIV prevalence if people migrate in or out of the study area based on their HIV status. Changes in HIV diagnostics could affect the timing of HIV diagnosis and/or how likely people are to get tested and, thus, could affect prevalence. Answer choice a is not correct because HIV is not a curable disease. HIV treatment may help an HIV-infected person live longer but will not affect whether or not they have HIV. Therefore, HIV prevalence is not affected by HIV treatment. Note that if a disease was curable with treatment then that disease's prevalence could be affected by treatment.

# ERIC Notebook

Second Edition

## Ecologic Studies


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Ecologic studies are studies in which the unit of observation is a group, not separate individuals, for one or more study variables. For example, exposure and risk factors are known only at the group level, such as the average air pollution concentration in different cities. The occurrence of the health outcome may also be only known at the group level, such as overall mortality rates from chronic lung disease in the same cities with measured levels of air pollution. Ecologic studies may be used to generate hypotheses of an association between exposure and a health outcome, but these studies cannot confirm causation. This is because, for example, we do not know whether those individuals who died in a particular city under observation had a higher exposure than individuals who remained alive.

### Ecologic fallacy

When dealing with group level information, it is important to be aware of what is called the ecologic fallacy. This fallacy results from concluding that because an association exists between exposure and a health outcome at the group level, it therefore exists at the individual level. The cause of this fallacy is that we do not know the link between exposure and the health outcome among individuals within each group. For example, we don't know the number of diseased persons who were exposed or not exposed in the high exposure group or in the low exposure group.

### Description of the ecologic fallacy

In ecologic studies, only information on aggregate measures, such as the average exposure in City A and the death rate in City A can be known. At the individual level, however, we can, for example, determine the proportion of people who died within each of the categories of exposure (low or high).

Suppose air pollution is higher in Baltimore than in Tampa, but mortality from lung disease is lower in Baltimore than in Tampa. It would be fallacious to conclude that air pollution protects against lung disease deaths. It is possible that persons dying of lung disease in Tampa may have moved from cities with high air pollution or that another risk factor for lung disease – such as smoking – is more prevalent in Tampa than Baltimore. We do not know the cumulative exposures of cases and non-cases in either city. The heterogeneity of lifetime air pollution

exposure among individuals in each city makes the average exposure unrepresentative of the distribution of exposure among individuals in the population.

Using the previous example, in-migrants to Tampa had previously experienced higher levels of exposure to air pollution, thus causing the aggregate level of lung disease deaths to appear higher in Tampa. The aggregate level of air pollution in Tampa does not allow us to see that there are the varying levels of exposure between lifelong residents and in-migrants, this shows how the ecologic fallacy can occur.

**Examples of questions investigated by ecologic studies include:**

- Is the ranking of cities by air pollution levels associated with the ranking of cities by mortality from cardiovascular disease, adjusting for differences in average age, percent of the population below poverty level, and occupational structure?
- Have new car safety measures such as passenger air bags made a difference in motor vehicle fatality rates in areas with different laws over the same period of time?
- Has introduction of newer cars with safety features such as multiple airbags made a difference in motor vehicle fatality rates in areas with a different distribution of newer versus older cars, over the same period of time?
- Are daily variations in mortality in Boston related to daily variations in particle air pollution, adjusting for season of year and temperature?
- What are the long-term time trends (1950-2012) for mortality from the major cancers in the U.S., Canada, and Mexico?

**Advantages of ecologic studies**

Aggregate data on exposure and health outcomes are often publicly available in state and national databases, such as the US census or from the Center for Disease Control and Prevention. Agencies of the state and federal government collect considerable data reported at the aggregate level on the economy, the environment, and the health and wellbeing of the population. Data are regularly obtained on air quality, water quality, weather conditions, the size of the population, the status of the economy, and the health of the population through surveys such as the National Health Interview Survey, National Health and Nutrition Examination, and Behavioral Risk Factor Survey. Data on the vital status of the population are obtained via birth and death registries and cancer and birth defects registries. These publicly available records provide databases for linking health outcomes with characteristics of the population, the environment, and the economy at the aggregate level.

Aggregate level data can conveniently be obtained by researchers at a low cost and can be useful for evaluating the impact of community-level interventions. Examples of interventions that may be evaluated through ecologic study designs include fluoridation of water, seat belt laws, and mass media health campaigns. Aggregate level information can be compared before and after the intervention to determine the effects of an intervention at the community level.

In addition, minimal within-community differences between exposures may exist; however exposures may differ substantially between communities, cities, states, and countries. Examples of small within-community exposure differences but large between-community differences include:

- Quality of drinking water
- Concentration of certain air pollutants such as ozone and fine particles
- Average fat content of diet (larger differences between countries than between individuals within the same city)
- Cumulative exposure to sunlight (larger differences by latitude [north-south] of residence than among individuals at the same latitude)

Ecologic studies are also useful for studying the effect of short-term variations in exposure within the same community, such as the effect of temperature on mortality.

### Types of ecologic study designs

There are three main types of ecologic study designs: cross-sectional ecologic studies, time-trend ecologic studies, and solely descriptive ecologic studies.

*Cross-sectional ecologic studies* compare aggregate exposures and outcomes over the same time period. An example of this study design is an investigation comparing bladder cancer mortality rates in cities with surface drinking water sources that contain chlorine by-products compared to rates in cities with ground drinking water sources that contain little or no chlorine by-products.

*Time-trend ecologic studies* compare variations in aggregate exposures and outcomes over time within the same community. A study investigating whether hospital admissions for cardiac disease in Los Angeles increase on days when carbon monoxide levels are higher would be an example of this type of study.

*Solely descriptive ecologic studies* investigate disease or risk factor differences between communities at the same time, or within the same community over time. This type of study design would be used to investigate the following questions: What are the differences in lung cancer mortality among cities in North Carolina? What is the secular trend of lung cancer mortality between 1960 and 2010 for the entire state of North Carolina?

**Ecologic Study Designs at a Glance**

| Study type | Design | Time frame |
|---|---|---|
| Cross-sectional | Across communities | Same time period |
| Time-trend | Within the same community | Over time |
| Descriptive | Across communities or<br>Within the same community | At a point in time or<br>Over time |

### Limitations of ecologic studies

Ecologic studies are subject to numerous biases and limitations. Most notably, these study designs are subject to the ecologic fallacy, which occurs by inferring that associations at the aggregate level are true at the individual level. Ecologic studies are also more often subject to confounding bias than are individual risk studies. *Confounding* is a mixing of the effects of other risk factors with the exposure of interest. Confounding bias may occur in an ecologic study if the confounding factor is correlated with the background rate of disease (the disease rate among unexposed persons in each study community). *Cross-level bias* occurs when the confounding factor is associated with the background rate of disease differentially across groups. The ecologic fallacy may occur as a result of cross-level bias.

**Example**

Suppose the association between average fat consumption and breast cancer risk is examined across communities in the US. Certain communities may have a larger percentage of women with a genetic predisposition to breast cancer than other communities. Suppose these same communities containing large percentages of women with a genetic predisposition to breast cancer are also communities with a high per capita dietary fat consumption. The results of the study will show a strong correlation between average dietary fat consumption and breast cancer mortality. The association will be inflated due to the confounding factor, genetic predisposition to breast cancer. This bias occurred because the background rate of breast cancer incidence

Time-trend ecologic studies are further limited in that an investigator cannot be confident that exposure preceded the outcome. Migration into and out of communities can also bias the interpretation of ecologic results.

## Practice Questions

*Answers are at the end of this notebook*

1) Researchers study the community of one town in North Carolina over a 10 year period. They conduct an ecologic study and collect data on the rate of automobile accidents each year among teenagers and the percentage of teens in the town who report using their phones for texting each year. Based on their data, the researchers conclude that teenagers who report texting on their phones are more likely to be in a car accident. Which of the following are true about the researchers' conclusion? Choose all that apply.

a) The researchers' conclusion is valid

b) The researchers have incorrectly used group-level data to draw conclusions about individual teenagers

c) The researchers do not know if the teens involved in car accidents are the same teens who have reported texting on their phones, therefore, their conclusion is not valid

d) The researchers' conclusion does not account for what other factors, if any, are related to texting and related to being in a car accident

2) Reported cases of the flu are higher in city A than city B. Vaccination rates for the flu are lower in city A than city B. Which of the following are reasons why would it be incorrect to simply assume that higher vaccination in city B is what is causing city B to have fewer reported cases of the flu? Choose all that apply.

a) City A and city B may have different strains of the flu

b) City A and city B may have different proportions of people in their populations who are especially vulnerable to the flu (e.g. the elderly, children and pregnant women)

c) City A and city B may have differences in health care accessibility, leading to differences in testing for the flu and diagnosis of the flu

d) City A and city B may have different climates, leading to differences in how/where people come into contact with each other. This may affect flu transmission rates

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. Answer choices b, c and d are correct. In the example, the researchers have used overall data on car accidents for a town and overall data on the prevalence of teens who reported texting to draw a conclusion about individuals. The researchers actually do not have any data on whether the teens involved in car accidents were the same teens who reported texting. In addition, there may be other factors involved that explain the study results such as another confounding variable that is related to being in a car accident and related to reporting texting.

2. All answer choices are correct. City B may have a different strain of the flu circulating, which could affect transmission rates, morbidity/mortality and whether or not an infected person seeks medical care. There may be more people in city A who are especially vulnerable to the flu and these vulnerable groups may experience greater illness, leading to more visits to a health provider and more diagnoses. If health care accessibility varies between the 2 cities and/or knowledge of testing and diagnosis for the flu differs between the 2 cities, this may affect whether infected persons seek health care and may affect which tests they receive. Such variability may lead to the appearance of greater or fewer cases. Finally, the climate may influence social mixing patterns leading to greater virus transmission in the city where people spend more time together indoors.

# ERIC Notebook

Second Edition

## Incident vs. Prevalent Cases and Measures of Occurrence


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

To determine which factors impact health outcomes at the population level, epidemiologists employ a number of different study designs. These designs are used to examine the relationship between exposures (or determinants) and health outcomes. A health outcome may be a disease, condition, death, event or a change in health status or behavior. For example, in addition to diseases, we may study health events such as injuries or the occurrence of an "event" such as preterm birth. Persons who experience the outcome of interest are commonly referred to as cases.

One of the first things to consider when developing a study is whether you will measure prevalent or incident cases.

Prevalent cases are all individuals living with the outcome of interest within a specified timeframe, regardless of when that person was diagnosed or developed the health outcome.

**Example**

In a study of prevalent cases of diabetes with a one year time period, anyone who has diabetes during the one year study period would be counted as a case.

These prevalent cases would include both people who have diabetes at the outset of the study year as well as any who developed diabetes over the course of the study.

*Incident cases* are all individuals who change in status from non-disease to disease – or from one state of a health outcome to another – over a specific period of time. In other words, "incidence" refers to the occurrence of new cases.

**Example**

For example, in a study of incident cases of diabetes with a one year time period, only those who developed diabetes over the course of the one year study period are considered incident cases.

### Measures of frequency

*Prevalence*

Prevalence is the proportion of a population living with a specific health outcome within a specified time. It is the only measure of occurrence calculated with prevalent cases. To calculate prevalence, the number of prevalent cases (numerator) is divided by the total population at risk

(denominator.) The total population at risk denominator includes the prevalent cases. Prevalence is often reported as a percentage.

**Prevalence = Prevalent cases / Total population**

Depending on the type of prevalence being calculated, the denominator can be either an average of the population over time or a single measurement at a specific point of time.

Prevalence can either be calculated as a *point prevalence* or *period prevalence.* A point prevalence is calculated with data from one specific point in time, while a period prevalence is calculated over a range of time.

Prevalence is directly affected by the incidence and duration of the health outcome under study, which makes it a poor choice for diseases or outcomes with a short duration or high mortality rate.

**Example**

For example, *Vibrio vulnificus* – a disease caused by consumption of raw shellfish – has a low incidence and short duration. Therefore, the few new (incident) cases that arise will remain prevalent in the population for only a short time before the cases recover or die. However, for a disease like diabetes, which has a higher risk or rate and longer duration, the prevalence will be higher than the risk or rate and is a valuable measure of the burden of disease in the population.

**Prevalence = Rate x Duration**

*Risk*

Like prevalence, risk is also a measure of the extent of a health outcome in a population. However, unlike prevalence, risk is the proportion of an at-risk population that develops a specific health outcome within a specified amount of time. The numerator for risk is incident cases, and the denominator includes only those at-risk of developing the outcome of interest at the beginning of study follow-up.

**Example**

If the disease under study is ovarian cancer, which obviously only affects women, the denominator should consist only of women in the population who, at the start of study observation, do not have ovarian cancer and are capable of developing ovarian cancer.

**Risk = Incident cases / Population at-risk**

Risks are often reported as a scaled value, such as cases per 1,000; 10,000; or 100,000 population.

*Rate*

Rate is another measure of health outcome occurrence calculated with incident cases of the health outcome. However, the denominator for a rate is the total amount of person-time at-risk. Person-time is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a study. In its simplest form, person-time is a sum of each study participant's time at risk before experiencing the outcome of interest or exiting the study. This is a better estimate of the true at-risk population because it excludes time for participants who are no longer eligible to experience the outcome of interest. Thus, rates are a better reflection of health outcome occurrence in a dynamic population, where participants may exit the study or become no longer at-risk.

**Rate = Incident cases / Total person-time at-risk**

The unit for a rate is "cases per person-time." Rates are often reported as a scaled value with a time unit relevant for the study, such as cases per 1,000 person-years, 12,000 person-months, or 365,000 person-days.

Rates are favored if the rapidity with which new cases of the health outcome or new events are occurring in the population is of interest.

### Comparing Measures of Occurrence

The image of the bathtub below graphically represents the relationships between prevalence, risk, and rate.

In this analogy, prevalence is the proportion of the tub (the total population) filled with any water (prevalent cases and incident cases). Risk is the proportion of the tub filled with new, flowing water (incident cases). Rate is a measure of how quickly the water flows into the tub. Prevalent cases only leave the prevalence pool by either recovery, death, migration out of the population or loss of study follow-up via the bathtub drain.

### Practice Questions

*Answers are located at the end of this notebook.*

1) If researchers were studying the risk of a woman having a baby born preterm in the United States in 2013, what would the at-risk population be? Assume for this question that a pre-term birth is any birth before 39 weeks gestation. Choose the one best answer.

a) All women in the United States in 2013
b) All pregnant women in the United States in 2013
c) The actual number of babies born preterm in the United States in 2013
d) All babies born in the United States in 2013
e) All pregnant women in the United States whose due dates would mean that the baby could potentially be born pre-term in the year 2013

2) Now researchers want to study the <u>risk</u> of head injuries that occur while a person was riding a bicycle, in the United States between 2000-2013. How would the risk denominator ideally be calculated? Choose the one <u>best</u> answer.

a) All people living in the United States in the time period 2000-2013
b) All people who rode a bicycle in the United States in the time period 2000-2013
c) All people who owned or had access to a bicycle in the United States in the time period 2000-2013
d) The actual number of documented head injuries due to bicycle accidents in the United States in the time period 2000-2013
e) All people who had a head injury in the United States in the time period 2000-2013

### Terminology

*Prevalent cases* – all individuals living with the health outcome of interest within a specified timeframe, regardless of when that person was diagnosed or developed the health outcome

*Incident cases* – all individuals who change in status from one state of health to another (such as non-disease to disease) over a specific period of time

*Prevalence* – the proportion of a population living with a specific health outcome within a specified timeframe

*Risk* – the proportion of an at-risk population that develops a specific health outcome within a specified amount of time

*Rate* – the frequency of incident cases per unit of person-time

### References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

Answers to Practice Questions

1. The best answer was e: All pregnant women in the United States whose pregnancy due date would mean that the baby could potentially be born pre-term in the year 2013. Since the researchers are studying risk, they need to measure incident cases. The denominator must include only those at-risk of having the outcome of interest during the defined time-period of interest (the year 2013).

Answer choice a (All women in the United States in 2013) is not the best choice because only pregnant women—and not all women in general—are at-risk of having a preterm baby born. Answer choice b (All pregnant women in the United States in 2013) is a good choice but is not the best answer choice among those listed. This is because a women who is pregnant in 2012 may be at-risk of giving birth to a pre-term baby is the calendar year 2013. So the researchers would not want to limit their at-risk population just to women who are pregnant in the United States in 2013. Answer choice c (The actual number of babies born preterm in the United States in 2013) is incorrect because this is not the at-risk population needed for a risk measure since this is a count of babies already born preterm. Similarly, answer choice d (All babies born in the United States in 2013) is incorrect because it does not represent the at-risk population that would need to be followed over time.

2. The best answer was b: All people who rode a bicycle in the United States in 2013. In your risk denominator, you want to only include people actually at-risk of the outcome. In order to have a head injury that arises from riding a bicycle, a person would actually have to be riding a bicycle, otherwise they would never be at-risk for the outcome under study.

Answer a (All people living in the United States in the time period 2000-2013) is not the best choice because all of these people did not ride a bicycle. Answer c (All people who owned or had access to a bicycle in the United States in the time period 2000-2013) is incorrect for the same reason. Just because someone has access to a bicycle does not mean they actually rode one. Answer d (The actual number of documented head injuries due to bicycle accidents in the United States in the time period 2000-2013) is not correct because the researchers are studying incident cases among all people at risk for the outcome. The group at-risk is all people who rode a bicycle and not just those who actually had a head injury. Answer e (All people who had a head injury in the United States in the time period 2000-2013) is incorrect because the researchers were not studying head injuries in general, they were studying head injuries arising from riding a bicycle specifically.

# ERIC Notebook

Second Edition

## Sources of Systematic Error or Bias: Information Bias


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS

Information bias is one type of systematic error that can occur in epidemiologic studies. *Bias* is any systematic error in an epidemiologic study that results in an incorrect estimate of the association between exposure and the health outcome. Bias occurs when an estimated association (risk ratio, rate ratio, odds ratio, difference in means, etc.) deviates from the true measure of association.

Bias is caused by systematic variation, while chance is caused by random variation. The consequence of bias is systematic error in the risk ratio, rate ratio, or odds ratio estimate. Bias may be introduced at the design or analysis phase of a study. We should try to eliminate or minimize bias through study design and conduct.

Major types of systematic error include the following:

- Selection bias
- Confounding bias
- Information bias

In this issue we present information bias. Selection bias and confounding are covered in separate ERIC Notebooks.

*Information bias* is a distortion in the measure of association caused by a lack of accurate measurements of key study variables. Information bias, also called measurement bias, arises when key study variables (exposure, health outcome, or confounders) are inaccurately measured or classified. Bias in the risk ratio, rate ratio, or odds ratio can be produced even if measured errors are equal between exposed and unexposed or between study participants that have or do not have the health outcome.

### Non-differential misclassification

Non-differential misclassification occurs if there is equal misclassification of exposure between subjects that have or do not have the health outcome or if there is equal misclassification of the health outcome between exposed and unexposed subjects. If exposure or the health outcome is dichotomous, then non-differential misclassification causes a bias of the risk ratio, rate ratio, or odds ratio towards the null.

### Non-differential misclassification of exposure status

Non-differential misclassification of exposure status in a case-control study occurs when exposure status is

equally misclassified among cases and controls. Non-differential misclassification in a cohort study occurs when exposure status is equally misclassified among persons who develop and persons who do not develop the health outcome.

Non-differential misclassification of health outcome status occurs in a case-control study when the health outcome status is equally misclassified among exposed and unexposed subjects. Non-differential misclassification of the health outcome status occurs in a cohort study when a study subject who develops the health outcome is equally misclassified among exposed and unexposed cohorts.

### Effect of non-differential misclassification of exposure

Non-differential misclassification biases the risk ratio, rate ratio, or odds ratio towards the null if the exposure classification is dichotomous, i.e., either exposed or unexposed. If exposure is classified into 3 or more categories, intermediate exposure groups may be biased away from the null, but the overall exposure-response trend will usually be biased towards the null.

### Effect of non-differential misclassification of the health outcome

In most cases, non-differential misclassification of the health outcome will produce bias toward the null, i.e. the risk ratio, rate ratio or odds ratio will be biased towards 1.0. If errors in detecting the presence of the health outcome are equal between exposed and unexposed subjects (i.e. sensitivity is less than 100%) but no errors are made in the classification of health outcome status (i.e. specificity is 100%), the risk ratio or rate ratio in a cohort study will not be biased, but the risk difference will be biased towards the null.

### Effect of non-differential misclassification of health outcome status

If no errors are made in detecting the presence of the health outcome (i.e. 100% sensitivity), but equal errors are made among exposed and unexposed in the classification of health outcome status (i.e. specificity less than 100%), the risk ratio, rate ratio, and risk difference (as applicable) will be biased towards the null.

Combined errors in both sensitivity and specificity further increase the bias towards the null, but specificity errors produce larger biases overall.

### Differential misclassification

Differential misclassification occurs when misclassification of exposure is not equal between subjects that have or do not have the health outcome, or when misclassification of the health outcome is not equal between exposed and unexposed subjects.

Differential misclassification causes a bias in the risk ratio, rate ratio, or odds ratio either towards or away from the null, depending on the proportions of subjects misclassified.

### Effect of differential misclassification of exposure or health outcome

Differential misclassification of the exposure or health outcome can bias the risk ratio, rate ratio, or odds ratio either towards or away from the null. The direction of bias is towards the null if fewer cases are considered to be exposed or if fewer exposed are considered to have the health outcome. The direction of bias is away from the null if more cases are considered to be exposed or if more exposed are considered to have the health outcome.

The effect of differential misclassification of the exposure or health outcome can bias the risk ratio, rate ratio, or odds ratio in either direction. The direction of bias is towards null if fewer cases are considered to be exposed or if fewer exposed subjects are considered to have the health outcome. The direction of bias is away from the null if more cases are considered to be exposed or if more exposed subjects are considered to have the health outcome.

### Interviewer bias

Interviewer bias is a form of information bias due to:

1. lack of equal probing for exposure history between cases and controls (exposure suspicion bias); or

2. lack of equal measurement of health outcome status between exposed and unexposed (diagnostic suspicion bias)

Solutions:

1. blind data collectors regarding exposure or health outcome status
2. develop well standardized data collection protocols
3. train interviewers to obtain data in a standardized manner
4. seek same information about exposure from two different sources, e.g. index subject and spouse in case-control study

**Recall or reporting bias**

Recall or reporting bias is another form of information bias due to differences in accuracy of recall between cases and non-cases or of differential reporting of a health outcome between exposed and unexposed.

Cases may have greater incentive, due to their health concerns, to recall past exposures. Exposed persons in a cohort study may be concerned about their exposure and may over-report or more accurately report the occurrence of symptoms or the health outcome.

Solutions:

1. add a case group unlikely to be related to exposure
2. add measures of symptoms or health outcomes unlikely to be related to exposure

**Complications in predicting direction of misclassification bias**

Misclassification of confounders results in unpredictable direction of bias. Non-differential misclassification of a polychotomous exposure variable (3 or more categories) may result in bias away from null, though this is less likely than bias towards the null.

Non-differential misclassification of a health outcome limited to a loss of sensitivity of detecting the health outcome without any loss in specificity does not bias toward null, whereas a loss of specificity always biases toward the null.

**Conclusions**

Some inaccuracies of measurement of exposure and health outcome occur in all studies.

If a positive exposure-health outcome association is found and non-differential measurement errors are more likely than differential ones, measurement error itself cannot account for the positive finding since non-differential error nearly always biases towards the null.

Strive to reduce errors in measurement:

1. develop well standardized protocols
2. train interviewers and technicians well
3. perform pilot studies to identify problems with questionnaires and measuring instruments
4. attempt to assess the direction of bias by considering likelihood of non-differential or differential misclassification

### Terminology

*Information bias:* A distortion in the measure of association caused by a lack of accurate measurements of exposure or health outcome status which can result from poor interviewing techniques or differing levels of recall by participants.

*Non-differential misclassification:* Equal misclassification of exposure between subjects that have or do not have the health outcome, or equal misclassification of the health outcome between exposed and unexposed subjects.

*Differential misclassification*: Unequal misclassification of exposure between subjects that have or do not have the health outcome, or unequal misclassification of the health outcome between exposed and unexposed subjects.

### Practice Questions

*Answers are at the end of this notebook*

1) Researchers conduct a case-control study. The following table shows the true classification of exposure and the health outcome (Note: these data are hypothetical and typically researchers would not know the true unbiased distribution of exposure and outcome).

| | Have health outcome | Do not have health outcome |
|---|---|---|
| Exposed | 200 | 210 |
| Unexposed | 340 | 500 |

a) Calculate the odds ratio

Now imagine that 50 people with the health outcome were misclassified as being unexposed and 20 people with the health outcome were misclassified as being exposed.

b) Create the corrected 2x2 table

| | Have health outcome | Do not have health outcome |
|---|---|---|
| Exposed | | |
| Unexposed | | |

c) Calculate the odds ratio for the corrected table
d) In which direction was the misclassification bias?

2) Researchers conduct a case-control study of the association between the diet of young children and diagnosis of childhood cancer, by age 5 years. The researchers are worried about the potential for recall bias since parents are being asked to recall what their children generally ate, over a period of 5 years. Which of the following potential control groups would be most likely to reduce the likelihood of recall bias?

a) Parents of children with no known health problems
b) Parents of children with other known, diagnosed serious health problems (aside from childhood cancer)
c) Parents of children with other known, diagnosed minor health problems

### References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

Dr. Steve Marshall and Dr. Jim Thomas Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, The University of North Carolina at Chapel Hill, Department of Epidemiology

Dr. David Richardson Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013, The

### Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of

Answers to Practice Questions

1.

a) Calculate the odds ratio

Odds ratio = (200*500) / (340*210) =1.4

b) Create the corrected 2x2 table

| | Have health out-come | Do not have health outcome |
|---|---|---|
| Exposed | 230 | 210 |
| Unex-posed | 310 | 500 |

c) Calculate the odds ratio for the corrected table

Odds ratio = (230*500) / (310*210) =1.8

d) In which direction was the misclassification bias?

The bias was away from the null (the null value is 1.0).

2. Answer choice b is the best choice. Researchers should aim to have similar recall bias between both the case and control groups. Parents of children who have childhood cancer, which is a serious health problem, are likely to be quite concerned about what may have contributed to the cancer. Thus, if asked by researchers, these parents are likely to think very hard about what their child ate or did not eat in their first years of life. Parents of children with other serious health problems (aside from cancer) are also likely to be quite concerned about any exposure that researchers ask about. Therefore, these parents can be expected to recall exposures in a way that is more comparable with parents of children who have cancer. In contrast, parents of children who have no health problems or parents of children with only minor health problems are less likely to be concerned with carefully recalling any exposures.

# ERIC Notebook

Second Edition

## Randomized Controlled Trials (Experimental Studies)


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


### What are randomized trials?

Randomized trials are epidemiological studies in which a direct comparison is made between two or more treatment groups, one of which serves as a control for the other. Study subjects are randomly allocated into the differing treatment groups, and all groups are followed over time to observe the effect of the different treatments. The control group may either be untreated (*placebo-controlled*) or undergo a "gold standard" established regimen against which the new regimen will be assessed (*active-controlled*). Randomized trials provide the most direct evidence for causality. However, they are also fraught with a number of additional considerations not present for observational research.

For example, unless researchers are genuinely uncertain about the potential harms or benefits of a treatment, it is unethical to assign it to one group of people while withholding it from others (*equipoise*). This limits the types of questions that can be answered using experimental studies.

A *placebo-controlled* randomized trial might compare the effect of vitamin E supplement in one group of schizophrenia patients (the treatment group) against the effects of a placebo on a separate group of schizophrenia patients (the control group).

An *active-controlled* randomized trial might compare diabetic patients with implanted insulin pumps against diabetic patients who receive multiple insulin injections (the control group).

Randomized Trial Study Design

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

### Randomization

Randomization avoids bias by eliminating baseline differences in risk between treatment and control groups. Randomization, if done properly, should make both groups similar in terms of the distribution of risk factors, regardless of whether these risk factors are known or unknown (thus eliminating confounding due to both measured and unmeasured variables). The larger the randomized groups, the greater the probability of equal baseline risks. However, participants in RCTs are often not representative of the target population, which introduces selection bias and limits generalizability.

### Methods of randomization

There are different ways to randomize study participants into treatment groups. A simple way to randomize would be to roll a die or use a random number table to allocate individuals into the different groups. Another way investigators randomize study participants is through stratified random allocation. Under this method, the investigator first stratifies the participants by a baseline risk factor (i.e., smoking status) then randomizes the subjects in each stratum into either the treatment or control group. Stratified random allocation is appropriate when the investigator wants to be sure that a strong external risk factor is equalized at baseline between treatment and control groups.

### Types of randomized trials

The two general types of randomized trials are *clinical trials* and *community trials*, with randomized clinical trials being by far the more common. A randomized clinical trial is an experiment with patients as subjects. The goal is to find an effective treatment for a disease or to evaluate an intervention to prevent the progression of a disease. Randomized clinical trials are often used to evaluate the efficacy of new drugs against standard treatments or against placebos, but they are also used to evaluate other therapeutic procedures such as a new form of surgery, a dietary regimen, or an exercise program for persons with pre-existing disease. Most often, patients who already have some specific disease are the subjects of study in clinical trials. However, at times, subjects who are at high risk for a specific disease are entered into a randomized clinical trial to assess the efficacy of a drug to prevent the disease. For example, women with a family history of breast cancer may be entered into a clinical trial to study the effect of the drug tamoxifen on the prevention of breast cancer.

A community trial is also an experiment, but differs from clinical trials in that an entire community, rather than an individual patient, is the unit of observation. For example, water fluoridation was evaluated by experimentally assigning entire communities to have their public water supply fluoridated or not fluoridated. Units of observation for a community trial may be a town or city, a factory or office, a classroom or an entire school. All persons in the same unit of observation are experimentally exposed to the same intervention although it is not certain that all persons in the unit will be equally exposed, e.g. that they will drink the fluoridated water coming from their taps. Several community trials have been conducted to evaluate the effectiveness of mass media campaigns to prevent heart disease by encouraging more exercise, less use of tobacco products, and other lifestyle modifications.

### Blinding or masking

Sometimes in clinical trials, participants, statisticians, and even investigators, are made unaware of whether the participants are part of the treatment or control group. When only study participants are unaware of their treatment status, but investigators and analysts are aware of treatment status, the trial is called *single-blinded*. When both the participants and the investigators are blinded as to the treatment status of the participants the trial is termed *double-blinded*. A *triple-blinded* trial is when subjects, investigators, and independent statisticians are kept unaware of subject treatment status.

Blinding the study participants by using placebos, or a sham treatment, is common practice in clinical trials. The

*placebo effect* occurs when participants report a favorable response when no treatment, but only placebo, is administered. Another bias that is prevented by blinding of subjects is *post-randomization confounding bias* where subjects' awareness of intervention may motivate them to be more cooperative or otherwise change their behavior. This motivation may correlate with other risk factors for the intended effect, thus destroying the design advantage of randomization.

**Example**

If individuals participating in a clinical trial to study the efficacy of a new weight loss drug are aware that they are receiving the weight loss drug, they may more closely comply with the prescribed study diet.

Another bias that is controlled for by blinding the subjects as to their treatment status is *selection bias*, or group differences in loss to follow-up. Symptoms of disease or side effects of the treatment may influence rates of loss to follow-up in subjects aware of their treatment status.

Bias due to differences in reporting of symptoms, a type of *information bias,* is also controlled by a double-blinded study. Study subjects who are aware of their treatment status may differentially report symptoms or side effects. Likewise, staff or statisticians may differentially evaluate subjects if they are aware of treatment status.

**Example**

In a study of the effects of a new drug on severity of migraines in which study members know their treatment status, the treated study members may believe that the drug will work and, therefore, report less severe migraines. If the investigator in this study knows the treatment status of the subjects, then that investigator may scrutinize the severity of the migraines in treated subjects more than that of the untreated subjects.

**Additional threats to the validity of a randomized trial**

Limiting the analysis to compliant subjects can create bias if compliance is correlated with other risk factors for the treatment effect. Analyzing the results without regard to subject compliance (called "intention-to-treat" analysis) can help to avoid this bias. That is, subjects should be included in the analysis whether or not they adhered to their treatment (or control) regimen.

**Example**

Suppose that in a clinical trial to look at the relationship between diet and risk of cancer, subjects were randomized to either a cancer-prevention diet or to a placebo diet. Suppose again that in the treatment group, those subjects with gastrointestinal symptoms that are precursors of cancer, were less compliant with their diet than subjects without symptoms. Exclusion from the analysis of the less compliant subjects would bias the results towards reporting a greater effect of the cancer-prevention diet. Only those subjects who were not at risk or who were at low risk of cancer would be included in the analysis. The appropriate analysis should include all persons originally assigned to their treatment group, whether or not they adhered to the treatments.

When noncompliant subjects are selectively excluded from an analysis, the benefit of randomization is lost, because unmeasured confounding factors may be associated with the lack of compliance.

**Treatment crossover**

Crossover, either planned or unplanned can create biases in experiments. In a *planned crossover*, group A (subjects treated with the new drug) and group B (subjects treated with a standard drug) would be switched to the other treatment at the midpoint of the trial. Two of the problems experienced with this experimental design are *carryover effects* and *diminished interest*. Carryover effects occur when the effects of the first drug last into the second half

of the study when the subjects are receiving the other treatment. Bias may also occur if there is diminished interest or lack of compliance in the second half of the study.

*Unplanned crossovers* occur when a clinician decides to switch a study member from the control to the treatment group, or vice versa, e.g. surgery vs. medical treatment for coronary artery disease. An unplanned crossover negates the benefit of randomization and introduces bias if switching is related to risk of the outcome.

### Loss to follow-up

Neither randomization nor blinding can prevent *differential loss to follow-up*, or more subjects dropping out in one treatment group than in another. Bias is introduced if the rate of loss to follow-up is correlated with both exposure to the treatment and exposure to other risk factors for the outcome.

**Threats to validity:**

Loss to follow-up
Non-compliance
Crossovers

### Analysis strategies to avoid bias

For purposes of analysis, study subjects should be kept in the original randomized group, even if they were lost to follow-up, switched to the other treatment group, or were non-compliant (the "intention-to-treat" principle). Analysis of any non-random subgroups threatens the validity of the study.

### Practice Questions

*Answers are located at end of this notebook.*

1) Researchers conducted a multi-year ongoing randomized controlled trial of the association between daily meditation (such as relaxation techniques) and health behavior among patients following a skin cancer diagnosis. Researchers randomly allocated study participants into 2 groups. The first patient group received weekly classes on meditation practices as well as a self-taught manual on meditation. The second patient group received only the self-taught manual on meditation. Data were collected at set intervals following the intervention to assess the patients' health behaviors. Studied health behaviors included data on the patient's diet, exercise, and mental health. Note: this is a hypothetical example.

a) In this example, why was it ethical for the researchers to allocate one group to receive weekly classes on meditation practices as well as a self-taught manual on meditation while the other group received only the self-taught manual on meditation?

b) Which of the following may bias the analysis? Choose all that apply.

a) Changes over time in how the health behaviors were defined and assessed

b) Inability to blind the researchers regarding which of the 2 meditation interventions each patient received

c) Lack of use of a separate untreated control group (e.g. a group that received no meditation intervention at all)

d) Patients that are not compliant with their assigned group (e.g. patients assigned to just the self-taught manual but who really want to be as healthy as possible so they show up at the weekly classes on meditation practices)

c) In this example study, would stratified random allocation have been useful?

### References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

### Answers to Practice Questions

1) In this example, researchers must be genuinely uncertain about the potential benefits of mediation in order for it to be ethical to assign different meditation interventions to different groups. If the effect of meditation on health behavior has not been extensively studied in this study population, then that would make it ethical to conduct this study.

b) Answer choices a, b, and d are correct. If researchers make changes in how they assess health behaviors over time, that would bias the results of the study. If researchers are not able to be blinded in regard to which patients are in which intervention group, this may lead researchers to assess or question the study participants about their health behaviors in different ways depending on which group they are in. Randomization works to eliminate baseline differences in risk between the 2 groups being compared. If some patients are not compliant with the group they were randomized to, this can negate the benefits of randomization. Lack of use of a separate untreated control group does not introduce bias into the study. The comparison groups are chosen based on the researchers' study hypothesis and based on what is ethical. If the researchers wanted to study the effect of weekly meditation classes + a self-taught manual on mediation versus just the self-taught manual on meditation then this is a valid study hypothesis.

c) Stratified random allocation may have been useful. Stratified random allocation is when the researchers first stratify participants by a baseline risk factor and then randomize the subjects in each stratum into the 2 comparison groups. Stratified random allocation can be done when the researchers want to be certain that a strong external risk factor is equalized at baseline between the 2 comparison groups. In our hypothetical example, researchers may have hypothesized that those participants who had strong family support would be more likely to learn and adopt meditation practices and more likely to have positive health behaviors. So the researchers could have first stratified participants based on the level of family support they reported and then, after that, randomized subjects into the 2 comparison groups.

### Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

# ERIC Notebook

Second Edition

## Risk and Rate Measures in Cohort Studies


**Second Edition Authors:**

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


Cohort studies are longitudinal studies where an exposed and an unexposed group (or less exposed group) are followed forward in time to find the incidence of the outcome of interest (e.g. disease, death, or change in health). Two measures of incidence are risks and rates. Risks and rates can be

further manipulated to provide additional information on the effects of the exposure of interest, such as risk or rate ratios, risk or rate differences, and attributable risk fractions.

Risk is defined as the number of new cases divided by the total population-at-risk at the beginning of the follow-up period. An individual's risk of developing the outcome of interest is

$$\text{Risk} = \frac{\#\ of\ new\ cases}{total\ \#\ individuals\ at\ risk}$$

measured.

A rate is the number of new cases of a health outcome divided by the total person-time-at-risk for the population. Person-time is calculated by the sum total of time all individuals remain in the study without developing the outcome of interest (the total amount of time that the study members are at risk of developing the outcome of interest). Person-time can be measured in days, months, or years, depending on the unit of time that is relevant to the study. A rate measures the rapidity of health outcome occurrence in the population.

$$\text{Rate} = \frac{\#\ of\ new\ cases}{total\ person\text{-}time\ at\ risk}$$

Two-by-two tables are generally used to organize the data from a study as shown below.

| | Disease | No disease | Total |
|---|---|---|---|
| Exposed | a | b | a+b |
| Unex-posed | c | d | c+d |
| Total | a+c | b+d | a+b+c+d |

**Risk ratios**

When risks are computed in a study, the risk ratio is the measure that compares the $\text{Risk}_{exposed}$ to the $\text{Risk}_{unexposed}$. The risk ratio is defined as the risk in the exposed cohort (the index group) divided by the risk in the unexposed cohort (the reference group). A risk ratio may vary from zero to infinity.

Risk Ratio = [a / (a+b)] / [c / (c+d)]

$$\text{Risk Ratio} = Risk_{exposed}/Risk_{unexposed}$$

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

For example, suppose researchers conduct a cohort study and gather the following data on the effects of gasoline fume exposure on respiratory illness among automotive workers.

| | Disease | No disease | Total |
|---|---|---|---|
| Exposed | 60 | 140 | 200 |
| Unex-posed | 25 | 175 | 200 |
| Total | 85 | 315 | 400 |

In this study, the risk in the exposed group is 60/200, or 0.30 cases per person (30 cases per 100 people), and the risk in the unexposed group is 25/200, or 0.125 cases per person (13 cases per 100 people). Therefore, the risk ratio is 0.30/0.125, or 2.4. A risk ratio of 2.4 implies that the exposed group has 2.4 times the risk of developing respiratory illness as the unexposed group.

### Rate ratio

When rates are computed in a study, the rate ratio is the measure that compares the $Rate_{exposed}$ to the $Rate_{unexposed}$. The rate ratio is defined as the rate of health outcome occurrence in the exposed cohort (the index group) divided by the rate of health outcome occurrence in the unexposed or less-exposed cohort (the reference group).

$$\text{Rate Ratio} = Rate_{exposed}/Rate_{unexposed}$$

A rate ratio measure also may show whether the exposure was preventive, harmful, or had no effect on the rate of health outcome in the exposed.

If in the previous example, the person-time-at- risk that each automotive worker contributed to the study had been recorded then the table might have looked like the following:

| | Disease | No disease | Person-years at risk |
|---|---|---|---|
| Exposed | 60 | 140 | 175 |
| Unexposed | 25 | 175 | 188 |
| Total | 85 | 315 | 363 |

In this study, the rate in the exposed cohort is 60/175 person- years, or 0.34 cases/person-year. The rate in the unexposed cohort is 25/188 person-years, or 0.13 cases/ person-year. The rate ratio in this study is 0.34/0.13, or 2.6, which is higher than the rate ratio calculated above. This rate ratio reveals that respiratory illness among workers exposed to gasoline fumes is developing at 2.6 times the rate that respiratory illness is developing among workers not exposed to gasoline fumes.

An exposure may be preventive (e.g., vitamin intake) or harmful (e.g., toxic chemical exposure). Confounding, which you will read about in another ERIC notebook issue, is one type of systematic error that can occur in epidemiologic studies. Confounding can cause an over- or under-estimate of the observed association between exposure and a health outcome. Assuming there are no other factors that may confound the association, a risk ratio less than 1 indicates that the risk in the exposed (index) group is less than the risk in the unexposed or less -exposed (reference) group, and therefore, the exposure is preventive. A risk ratio or rate ratio that equals 1 (the null value) indicates that there is no difference in risk or rates between exposed and unexposed groups. A risk ratio greater than one indicates that the risk in the exposed is greater than the risk in the unexposed, and, therefore, the exposure is harmful.

The following table may be applied to both risk and rate ratios.

| Risk ratio or | Exposure |
|---|---|
| <1 | Exposure is protective |
| =1 | Exposure is neither preventive nor harmful (null association) |
| >1 | Exposure is harmful |

The farther away the risk ratio or rate ratio is from the null value of one, the greater the effect of exposure is on the study group. This is shown in the following diagram.

RR=risk ratio or rate

## Risk and Rate differences

Difference measures are absolute measures of disease burden and helpful for public health program planning. Ratio measures are relative measures; they are commonly used to research the etiology of disease. Risk and rate differences answer the question "how much disease is attributed to the exposure?" Risk difference is defined as the risk in the exposed minus the risk in the unexposed (Equation 1). The risk difference is the *excess* risk of disease among the exposed population. Rate difference uses a similar equation (Equation 2.) In the past, risk difference was called "attributable risk"; sometimes attributable risk is still used. "Attributable fraction among the exposed" is the risk difference reported as a percent of the exposed population (Equation 3). "Attributable proportion" or "attributable risk percent" are alternative terms for Equation 3. Lastly, the attributable fraction among the total population (Equation 4) answers the question "what proportion of disease in the *total* population is associated with the exposure?" See figure, terms, and equations defined below.

### Figure

Adapted from Gordis L. Epidemiology, p. 156

**Equation 1:** Risk Difference= $\text{Risk}_{\text{exposed}} - \text{Risk}_{\text{unexposed}}$

**Equation 2:** Rate Difference = $\text{Rate}_{\text{exposed}} - \text{Rate}_{\text{unexposed}}$

**Equation 3:** Attributable fraction among the exposed:

$$\text{RD\%(exposed)} = \frac{\text{risk in exposed} - \text{risk in unexposed}}{\text{risk in exposed}} \times 100$$

**Equation 4:** Attributable fraction among the total population:

$$\text{RD\%(total population)} = \frac{\text{risk in total pop} - \text{risk in unexposed}}{} \times 100$$

### Practice Questions

*Answers are at the end of this notebook*

1) Researchers conduct a prospective cohort study to assess the association between dietary supplements and cognitive ability among children. A total of 500 children age 12-17 years who take an omega-3 fatty acid supplement are compared with 500 children age 12-17 years who do not take an omega-3 fatty acid supplement. Researchers follow the children for 2 years. During this time, 300 children who take the supplement earn what is classified as a "high" score on a cognitive test while 200 children who do not take the supplement earn what is classified as a "high" score on the same cognitive test.

a) Construct a 2x2 table from the information presented above

b) The risk difference is:

c) The attributable fraction among total population is:

d) The attributable fraction among the exposed is:

2) Researchers conduct a prospective cohort study of the association between working at a high-stress job and depression. A total of 5000 adults are followed for an average of 10 years. Among these 5000 adults, 2500 had a high-stress job while 2500 had a low-stress job. Participants had yearly health checks. A total of 250 incident cases of depression were diagnosed in the high-stress job group while 90 incident cases of depression were diagnosed in the low-stress job group. Assume all cases of depression were diagnosed at the end of year 5 of follow-up.

a) Construct a 2x2 table from the information above

b) Rate exposed =

Rate unexposed =

c) Rate ratio =

d) Rate difference=

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health. 1998 Jan;88(1):15-9. No abstract available. Erratum in: Am J Public Health. 2008 Dec;98(12):2119

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Center at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

## Answers to Practice Questions

1.

a)

| | Scored "high" on exam | Did not score "high" on exam | Total |
|---|---|---|---|
| Exposed to supplement | 300 | 200 | 500 |
| Unexposed to supplement | 200 | 300 | 500 |
| Total | 500 | 500 | 1000 |

b) Risk exposed = 300/500= 0.6 cases/person
Risk unexposed = 200/500 = 0.4 cases/person
Using equation 1, we calculate the Risk difference as 0.6 – 0.4 = 0.2 cases/person or 20 cases per 100

c) Attributable fraction in total population (equation 4)
Risk in Total population 500/1000 = 0.5 cases/person
Attributable risk in total population= (0.5-0.4)/0.5= 0.2 cases/person or 20 cases per 100

d) Attributable fraction in exposed (equation 3)
Risk exposed = 300/500= 0.6 cases/person
Risk unexposed = 200/500 = 0.4 cases/person
Attributable risk in exposed = (0.6-0.4)/0.6=0.33 cases/person or 33 cases per 100

2.

a)

| Study group | Depression | No depression | Total person-years |
|---|---|---|---|
| High-stress job | 250 | 2250 | (250*5+2250*10) = 23,750 |
| Low-stress job | 90 | 2410 | (90*5+2410*10)= 24,550 |
| Total | 340 | 4660 | 48, 300 |

b) Rate exposed = 250/23,750 person years * 10,000 = 105.3 cases/ 10,000 person years

Rate unexposed = 90/24,550 person years * 10,000 = 36.7 cases/ 10,000 person years

c) Rate ratio = 105.3 cases/ 10,000 person years / 36.7 cases/ 10,000 person years = 2.87

# ERIC Notebook

Second Edition

## Selection Bias


Second Edition Authors:

Lorraine K. Alexander, DrPH

Brettania Lopes, MPH

Kristen Ricchetti-Masterson, MSPH

Karin B. Yeatts, PhD, MS


Selection bias is a distortion in a measure of association (such as a risk ratio) due to a sample selection that does not accurately reflect the target population. Selection bias can occur when investigators use improper procedures for selecting a sample population, but it can also occur as a result of factors that influence continued participation of subjects in a study. In either case, the final study population is not representative of the target population – the overall population for which the measure of effect is being calculated and from which study members are selected.

Selection bias occurs when the association between exposure and health outcome is different for those who complete a study compared with those who are in the target population.

**Example**

In a case-control study of smoking and chronic lung disease, the association of exposure with disease will tend to be weaker if controls are selected from a hospital population (because smoking causes many diseases resulting in hospitalization) than if controls are selected from the community.

In this example, hospital controls do not represent the prevalence of exposure (smoking) in the community from which cases of chronic lung disease arise. The exposure-disease association has been distorted by selection of hospital controls.

**Sources of selection bias.**

*Selective survival and losses to follow-up*

After enrollment of subjects and collection of baseline data there is usually some loss to follow-up, i.e. when individuals leave the study before the end of follow-up. This biases the study when the association between a risk factor and a health outcome differs in dropouts compared with study participants.

*Volunteer and non-response bias*

Individuals who volunteer for a study may possess different characteristics than the average individual in the target population. Individuals who do not respond to requests to be studied generally have different baseline characteristics than responders. Bias will be introduced if the association between exposure and a health

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH

outcome differs between study volunteers and non-responders.

*Hospital patient bias (Berkson's Bias)*

Berkson's bias may occur when hospital controls are used in a case-control study. If the controls are hospitalized due to an exposure that is also related to the health outcome under study, then the measure of effect may be weakened, i.e. biased towards the null hypothesis of no association.

*Healthy worker effect*

Generally, working individuals are healthier than individuals who are not working. Therefore, in occupational exposure studies, where cases (or exposed subjects) are workers, controls (or unexposed subjects) should also be workers, otherwise the association between exposure and the health outcome will tend to be biased towards the null.

### Selection Bias

Selection bias will occur as a result of the procedure used to select study participants when the selection probabilities of exposed and unexposed cases and controls from the target population are differential and not proportional. This can occur when exposure status influences selection.

Selection bias will occur in cohort studies if the rates of participation or the rates of loss to follow-up differ by both exposure and health outcome status. Although we seldom can know the exposure and health outcome status of non-respondents or persons lost to follow-up, it is sometimes possible to obtain these data from an external source.

### Terminology

*Bias:* a systematic error in a study that leads to a distortion of the results. ( *Target population:* the overall population for which the measure of effect is being calculated, and from which study members are selected.

*Loss to follow-up:* when individuals leave the study before the end of follow-up.

Medical Epidemiology, Greenberg RS, 1993).

### Practice Questions

*Answers are at the end of this notebook*

1) Researchers are planning to conduct a case-control study of the association between an occupational exposure and a health outcome. The researchers plan to study exposed workers from one factory and compare them with unexposed retirees who have never worked in a factory. A reviewer of the research proposal is worried about selection bias and in particular about the possibility of the healthy worker effect. Which of the following best represents the reviewer's concern?

a) Retirees should not be compared to factory workers because factory workers are under more stress than retirees

b) Retirees should not be compared to factory workers because factory workers' incomes differ from those of retirees

c) Retirees should not be compared to factory workers because factory workers are likely to need to maintain a certain level of health in order to work in a factory while retirees would not necessarily be as healthy

d) Retirees should not be compared to factory workers because factory workers likely live in a different city than the retirees

2) Researchers conducted a prospective cohort study of the association between air pollution exposure and asthma. Some study participants were lost to follow-up (dropped out of the study) over time. The researchers were able to obtain data on the exposure and the health outcome for participants who remained in the study as well as for participants who dropped out of the study. The researchers discovered that the rate of loss to follow-up did not differ when comparing exposed and unexposed groups. The researchers also found that the rate of loss to follow-up did not differ when comparing people who developed asthma and people who did not develop

asthma. Based on this information, which one of the following statements is most likely to be true?

a) Selection bias likely occurred in this study because both exposure groups experienced loss to follow-up

b) Selection bias likely did not occur in this study because exposure status and health outcome status did not influence whether or not people dropped out of the study

c) Selection bias likely occurred in this study because both of the outcome groups (people with asthma and people without asthma) experienced loss to follow-up

d) Selection bias likely did not occur in this study because people cannot choose if they are exposed to air pollution or not exposed to air pollution

## References

Dr. Carl M. Shy, Epidemiology 160/600 Introduction to Epidemiology for Public Health course lectures, 1994-2001, The University of North Carolina at Chapel Hill, Department of Epidemiology

Rothman KJ, Greenland S. Modern Epidemiology. Second Edition. Philadelphia: Lippincott Williams and Wilkins, 1998.

The University of North Carolina at Chapel Hill, Department of Epidemiology Courses: Epidemiology 710, Fundamentals of Epidemiology course lectures, 2009-2013, and Epidemiology 718, Epidemiologic Analysis of Binary Data course lectures, 2009-2013.

## Acknowledgement

The authors of the Second Edition of the ERIC Notebook would like to acknowledge the authors of the ERIC Notebook, First Edition: Michel Ibrahim, MD, PhD, Lorraine Alexander, DrPH, Carl Shy, MD, DrPH, Gayle Shimokura, MSPH and Sherry Farr, GRA, Department of Epidemiology at the University of North Carolina at Chapel Hill. The First Edition of the ERIC Notebook was produced by the Educational Arm of the Epidemiologic Research and Information Cener at Durham, NC. The funding for the ERIC Notebook First Edition was provided by the Department of Veterans Affairs (DVA), Veterans Health Administration (VHA), Cooperative Studies Program (CSP) to promote the strategic growth of the epidemiologic capacity of the DVA.

## Answers to Practice Questions

1. Answer choice c is correct. The healthy worker effect is a type of selection bias that may occur in occupational exposure studies when the exposed cases are workers but the non-exposed study participants (controls) are not workers. In general, working individuals are healthier than non-working individuals. Health problems may actually be a reason for not working. In addition, retirees are typically older than the working population and may have more age-related health problems.
2. Answer choice b is correct. Selection bias likely did not occur in this study because exposure status did not influence whether or not people dropped out of the study. Furthermore, the health outcome status did not influence whether or not people dropped out of the study. Remember that selection bias may occur in a cohort study if the rate of participation or the rate of loss to follow-up differ by both exposure and health outcome status. Selection bias is not affected by if the exposure is an avoidable exposure or a non-avoidable exposure.

# Lecture 1: Introduction and Overview

*Linear models* are used to study how a quantitative response variable depends on one or more explanatory variables. The model is called *linear* because it assumes a linear relationship,

$$E(y|x) = \beta_0 + \beta_1 x,$$

between a response $y$ and predictor(s) $x$.

## WHY USE LINEAR MODELS?

- *A Scientific Question:* Suppose an obstetrician is interested in knowing whether babies of women who smoke during pregnancy weigh less than babies of women who do not smoke. (Low birth weight is associated with increased morbidity and mortality of infants.) The obstetrician has a dataset containing information about all the deliveries at County General Hospital. For each delivery, you know the birth weight of the infant and whether or not the mother smoked.

- *Your Initial Analysis:* To answer the scientific question of interest, you conduct a two sample t-test to compare birth weights of infants with smoking mothers to the birth weights of infants with nonsmoking mothers.

  *Question:* How does the t-test answer the question of interest?

- *Physician Reaction:* The obstetrician was thrilled with the results of your analysis (clearly, the p-value was $< 0.05$).
- *Physician Reconsideration:* On the way out of your office, the physician comments that teenaged mothers are more likely to have low birth weight infants than older mothers and are also more likely to smoke. In addition, the obstetrician comments that perhaps the number of cigarettes smoked per day is related to the birth weight as well. Do more cigarettes lead to lower birth weight? What about family income or maternal stress? Alcohol use or cocaine use? Smoking marijuana? What about women who stop smoking or reduce smoking levels after learning about the pregnancy?

- *Secondary Analysis:* You realize that the t-test does not control for these additional factors that may be related to birth weight and may influence the relationship between cigarette smoking and birth weight. Accurate estimation of the effect of cigarette smoking on birthweight will depend on how smoking, as well as other variables like maternal age and drug use, are related to birthweight and each other. You need to use a *model* that incorporates the various exposures and potential confounders into your analysis; such *multiple regression* models are extremely important in observational (i.e., non-randomized) studies.

Linear models are one of the world's most popular statistical tools. Understanding the theory of linear models is also a foundation for understanding more complex models used in statistics (including generalized linear models, longitudinal and multivariate models, and survival models).

## Connection to Previous Coursework

In BIOS662, we considered the basics of sampling, one- and two-sample parametric and nonparametric inference, and simple methods for the analysis of data from more than two groups (ANOVA and simple linear regression).

In BIOS663, we will discuss the general linear model in detail.

- We begin by discussing standard results for least-squares model fitting and basic procedures of inference (testing simple and complex hypotheses, constructing confidence intervals and regions, and making predictions).
- Then, we move to aspects of model checking, including residual analysis and detection and treatment of outliers.
- Next, we discuss basic procedures for inference in polynomial models

and consider basic smoothing techniques.

- We devote considerable time to model-building, variable selection, and model validation.
- We conclude by considering the generalized linear model (logistic and Poisson regression in particular) and methods for correlated response data.

## Example 1: One-Way ANOVA

Analysis of Variance (ANOVA) involves comparing random samples from several populations. One-way ANOVA is an extension of the 2-sample t-test to three or more samples.

*Hypothesis:* $NO_2$ exposure leads to damaged lung tissue in mice.

Lung damage was measured by percent serum fluorescence, where higher readings indicate greater damage. Investigators exposed 30 mice to a high dose of $NO_2$, 30 mice to a low dose, and selected 30 mice as controls. (Nitrogen dioxide is found in tobacco smoke and can also be produced by kerosene heaters and unvented gas stoves.)

Define:

$y_{ij}$= serum fluorescence of the $j^{th}$ mouse in the $i^{th}$ dose group, $i = 1, \ldots, 3$, $j = 1, \ldots, 30$.

$\mu_i$=average serum fluorescence in group $i$

We have the model:

$y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, \ldots, 3$, $j = 1, \ldots, 30$,

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$ is the vector of parameters to be estimated. We assume that the $\varepsilon_{ij}$'s are *i.i.d.* from some distribution with $E(\varepsilon_{ij}) = 0$ and Var$(\varepsilon_{ij}) = \sigma^2$. This model is often called a *means* model.

We may wish to use this model to do the following.

- Estimate the $\mu_i$'s and $\sigma^2$.
- Test hypotheses about the $\mu_i$'s, for example,

  $$H_0 : \mu_1 = \mu_2 = \mu_3.$$

  What does this hypothesis mean in terms of the subject matter?
- Decide which group mean is the largest, smallest, etc.

## Review of Terminology and Basic Concepts

### Scales of Measurement

Scales of measurement help to determine what type of analysis or model should be used for the data.

- A *nominal* variable is for mutually exclusive, but not ordered, categories, e.g. blood type or gender; also called *categorical*
- An *ordinal* variable is one where the order matters but not the difference between values, e.g. AP Basketball poll (also called *ranked data*) or pain severity scale (none, minor, moderate, severe)
- An *interval* variable is a measurement where the difference between two values is meaningful. The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees. Examples include temperature in degrees Centigrade.

- A *ratio* variable has all the properties of an interval variable, and also has a clear definition of 0. When the variable equals 0, there is none of that variable. Variables like height, weight, enzyme activity are ratio variables. Temperature, expressed in F or C, is not a ratio variable. A temperature of 0 on either of those scales does not mean 'no temperature'. However, temperature in degrees Kelvin is a ratio variable, as 0 degrees Kelvin really does mean 'no temperature'. When working with ratio variables, but not interval variables, you can look at the ratio of two measurements. A weight of 4 grams is twice a weight of 2 grams, because weight is a ratio variable. A temperature of 100 degrees C is not twice as hot as 50 degrees C, because temperature C is not a ratio variable.

Interval and ratio variables are types of *continuous* variables. Ratio variables typically have error variance proportional to the size of the measurement, so transformations are often used to stabilize the variance.

## Types of Variables

- *Response* or *Dependent Variable*: variable that is to be described in terms of other variables
- *Predictor* or *Independent Variable* or *Covariate*: used (perhaps with other independent variables) to describe a response variable
- *Control* or *Nuisance Variables*: may affect relationships but of no real interest in current study.
  - *Confounder*: a third factor that can lead to an observed association (or lack of association) due in fact to mixing of effects between the dependent variable, independent variable, and the confounding variable. For example, let the dependent variable be development of lung cancer, the independent variable be smoking, and the potential confounder be a genetic mutation. This particular genetic mutation may increase the likelihood of

lung cancer, as well as the likelihood to be addictive to smoking.

## Sets of Interest

- *Population:* any set of interest. A *parameter* describes a property of a population.
- *Sample:* any subset of a population. A *statistic* describes a property of a sample.

## Statistical Activities

- *Parameter Estimation:* means of providing a value (the "estimator") thought to be a "good" approximation of a parameter.

  What is "good"?

  - An estimate $\hat{\theta}$ of a parameter $\theta$ is *consistent* if $\hat{\theta} \to \theta$ as $n \to \infty$.
  - An estimate $\hat{\theta}$ of a parameter $\theta$ is *unbiased* if $E(\hat{\theta}) = \theta$.
  - *Maximum likelihood* estimates are asymptotically unbiased and consistent.

- *Inference:* Statistical inference or statistical induction comprises the use of statistics and random sampling to make inferences concerning some unknown aspect of a population. It is distinguished from descriptive statistics. Two schools of statistical inference are frequency probability and Bayesian inference.

## Purposes of Statistical Analysis

- *Confirmatory:* Relies on prespecified variables, model, and hypothesis test of interest. Generally, when pharmaceutical companies evaluate drugs in an effort to obtain FDA (U.S. Food and Drug Administration) approval, they conduct confirmatory analyses. (Does the new drug lead to a better patient outcome than the standard treatment?)

- *Exploratory:* Seeks to find patterns or explain variation in data; "fishing expedition." A major epidemiologic study, for example, might involve a primary confirmatory-type analysis for which a study is designed (Do infections during pregnancy lead to dangerously early births?) and several exploratory analyses (Do factors such as diet, exercise, socioeconomic status, education, drug use, etc. also affect the length of pregnancy?).

In an exploratory analysis, p-values generally must be interpreted with caution due to multiple testing of the data.

## General Linear Model (GLM)

**General:** Applicable to wide variety of problems of estimation and testing

**Linear:** Regression function, $E(y_i \mid \mathbf{x}_i)$ is a linear function of the parameters $\boldsymbol{\beta}$

**Model:** Describes the relationship between the one response and one or more predictors.

Muller & Fetterman terminology, also known as "Multiple Regression" in ALR. A simple *model function* is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with corresponding *regression function*

$$E(y_i|x_i) = \mu_i = \beta_0 + \beta_1 x_i.$$

The GLM is a *univariate* model, which means that we consider only one response variable. For this reason, the GLM is sometimes called the GLUM (general linear univariate model, or Simple Linear Regression in ALR). We may have one predictor or many predictors, in which case we may call our model a *multivariable* model, or refer to the regression problem as *mutiple regression*.

*Multivariate* models involve more than one response (researchers outside biostatistics sometimes use the word *multivariate* instead of *multivariable* to describe models with one response and many predictors).

## Example 2: Wingspan and Height Relationship

**Hypothesis:** Wingspan (the distance from one outstretched fingertip to the other) is equal to one's height.

$y_i$=wingspan of subject $i$, $i = 1, \ldots, n$

$x_i$=height of subject $i$

We have the model:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$.

Hypotheses to test: $H_0 : \beta_1 = 1$ and $\beta_0 = 0$.

## Example 3: UNC Faculty Salaries

Suppose we wish to model the relationship between faculty salary and years of service at UNC. We define the varables $y_i$, the salary of faculty member $i$, $i = 1, \ldots, n$, and $x_i$, the years of service at UNC for faculty member $i$, and fit the model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

which may be written in matrix form as

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}_{n\times 2} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}_{2\times 1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n\times 1} .$$

Now suppose we are interested in the relationship between rank (instructor, assistant professor, associate professor, or full professor) and faculty salary. We define three indicator variables, $x_1, x_2, x_3$, as follows:

$$x_1 = \begin{cases} 1 & \text{assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{associate professor} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{full professor} \\ 0 & \text{otherwise.} \end{cases}$$

In this coding scheme, instructors have $(x_1, x_2, x_3) = (0, 0, 0)$ (reference group), assistant professors have $(x_1, x_2, x_3) = (1, 0, 0)$, associate professors have $(x_1, x_2, x_3) = (0, 1, 0)$, and full professors have

$(x_1, x_2, x_3) = (0, 0, 1)$. An ANOVA model is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1,n_{\text{assistant}}} \\ y_{21} \\ \vdots \\ y_{2,n_{\text{associate}}} \\ y_{31} \\ \vdots \\ y_{3,n_{\text{full}}} \\ y_{41} \\ \vdots \\ y_{4,n_{\text{instructor}}} \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \end{pmatrix}_{n\times 4} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1,n_{\text{assistant}}} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2,n_{\text{associate}}} \\ \varepsilon_{31} \\ \vdots \\ \varepsilon_{3,n_{\text{full}}} \\ \varepsilon_{41} \\ \vdots \\ \varepsilon_{4,n_{\text{instructor}}} \end{pmatrix}_{n\times 1}.$$

Next, suppose that we are interested in evaluating both rank and years at UNC in one regression model. An ANCOVA model is given by

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1,n_{\text{asst}}} \\ y_{21} \\ \vdots \\ y_{2,n_{\text{assoc}}} \\ y_{31} \\ \vdots \\ y_{3,n_{\text{full}}} \\ y_{41} \\ \vdots \\ y_{4,n_{\text{inst}}} \end{pmatrix}_{n\times 1} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & x_{1,n_{\text{asst}}} \\ 1 & 0 & 1 & 0 & x_{21} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & x_{2,n_{\text{assoc}}} \\ 1 & 0 & 0 & 1 & x_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3,n_{\text{full}}} \\ 1 & 0 & 0 & 0 & x_{41} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & x_{4,n_{\text{inst}}} \end{pmatrix}_{n\times 5} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1,n_{\text{asst}}} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2,n_{\text{assoc}}} \\ \varepsilon_{31} \\ \vdots \\ \varepsilon_{3,n_{\text{full}}} \\ \varepsilon_{41} \\ \vdots \\ \varepsilon_{4,n_{\text{inst}}} \end{pmatrix}_{n\times 1}.$$

Finally, we may consider an interaction between rank and years of service in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} & x_{11} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & x_{1,n_{\text{asst}}} & x_{1,n_{\text{asst}}} & 0 & 0 \\ 1 & 0 & 1 & 0 & x_{21} & 0 & x_{21} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & x_{2,n_{\text{assoc}}} & 0 & x_{2,n_{\text{assoc}}} & 0 \\ 1 & 0 & 0 & 1 & x_{31} & 0 & 0 & x_{31} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 & x_{3,n_{\text{full}}} & 0 & 0 & x_{3,n_{\text{full}}} \\ 1 & 0 & 0 & 0 & x_{41} & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & x_{4,n_{\text{inst}}} & 0 & 0 & 0 \end{pmatrix}_{n \times 8}$$

This is *full model in every cell* by Muller and Fetterman.

## Next: Linear Algebra Review

*Reading Assignment*

- Weisberg, Appendix A.6-7: "A Brief Introduction to Matrices and Vectors"
- Muller and Fetterman, Appendix A: "Matrix Algebra Useful for Linear Models"
- Namboodiri: "Matrix Algebra: An Introduction" (Optional)

*To do*

- Download and install R: `https://cloud.r-project.org/`
- Order and install SAS: `https://software.sites.unc.edu/software/`

# Lecture 2: Linear Algebra Review

Reading

- Weisberg, Appendix A.6-7: "A Brief Introduction to Matrices and Vectors"
- Muller and Fetterman, Appendix A: "Matrix Algebra Useful for Linear Models"
- Namboodiri: "Matrix Algebra: An Introduction" (Optional)

## Why Linear Algebra?

Data for linear models can be can be represented as vectors and matrices

Simplification of theory and estimation of linear models, general representation

Easier to address certain problem in estimating such models

## Basics of Notation

A *matrix* is a two-dimensional array of elements.

$\mathbf{A} = \{a_{ij}\}$ means $\mathbf{A}$ is the matrix whose $i, j^{th}$ element (or component) is the *scalar* $a_{ij}$, where $i$ indexes row and $j$ indexes column, $i = 1, \ldots, r$, $j = 1, \ldots, c$. In this context, a scalar $s$ will represent a real number. Capital letters are used to represent matrices, and lowercase letters are used for vectors.

The *dimension* of $\mathbf{A}$ is $(r \times c)$ (say "r by c"), often written $\mathbf{A}_{r \times c}$.

The entire matrix $\mathbf{A}$ may be represented

$$\mathbf{A}_{r \times c} = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1c} \\ a_{21} & \ddots & & \ldots \\ \ldots & & \ddots & \ldots \\ a_{r1} & \ldots & \ldots & a_{rc} \end{bmatrix}.$$

A matrix with one column, i.e. an $(r \times 1)$ matrix, is called a *vector* or

*column vector*. A $(1 \times r)$ matrix is called a *row vector*. We can represent the matrix $\mathbf{A}_{r \times c}$ in terms of its columns $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_c\}$. For example, we might represent the $\mathbf{X}$ matrix from the one of the examples in lecture 1 as

$$\begin{aligned} \mathbf{X} &= \{\mathbf{1}, \textbf{height}\} \\ &= \begin{bmatrix} 1 & 61.61 \\ 1 & 59.17 \\ 1 & 54.75 \\ \vdots & \vdots \\ 1 & 62.92 \end{bmatrix}. \end{aligned}$$

## Types of Matrices

A *square matrix* has the same number of rows as columns, so that $r = c$.

For $\mathbf{A}_{r \times c}$ and $r \geq c$, the *diagonal* of $\mathbf{A}$ is $\{a_{11}, a_{22}, \ldots, a_{cc}\}$.

A *symmetric matrix* is a square matrix with $a_{ij} = a_{ji}$ for all $i, j$. That is, the entry in row $i$ and column $j$ is the same as the entry in row $j$ and column $i$. Only square matrices can be symmetric. The matrix

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix}$$

is symmetric (note that we refer to symmetry about the main diagonal, which runs from the upper left to the lower right).

A square matrix is called a *diagonal matrix* if all elements off the main diagonal are zero; that is, if $i \neq j$, then $a_{ij} = 0$. We write $\text{Diag}(\mathbf{b}) = \text{Dg}(\mathbf{b})$ for a square diagonal matrix created from a vector. For example,

$$\text{Diag}\left(\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}\right) = \begin{bmatrix} b_1 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & b_3 \end{bmatrix}.$$

An *identity matrix*, denoted $\mathbf{I} = \mathbf{I}_n = \mathbf{I}_{n \times n}$, is a diagonal matrix with all 1's on the main diagonal and 0's elsewhere. That is, $a_{ij} = 1$ for $i = j$ and $a_{ij} = 0$ for $i \neq j$. Thus

$$\mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Denote an $(n \times 1)$ vector of 1's by

$$\mathbf{1} \quad = \quad \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1} \quad = \mathbf{J}_n.$$

A *zero matrix* or *null matrix*, denoted $\mathbf{0}_{r \times c}$, has $a_{ij} = 0$ for all $i, j$. So

$$\mathbf{0_{2 \times 2}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

An *upper triangular matrix* $\mathbf{A}$ has $a_{ij} = 0$ for $i > j$. That is,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

is an upper triangular matrix. A *lower triangular matrix* has $a_{ij} = 0$ for $i < j$.

A *partitioned* matrix has elements grouped into submatrices by combinations of vertical and horizontal slicing. For example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \vdots & a_{13} \\ a_{21} & a_{22} & \vdots & a_{23} \\ \dots & \dots & \dots & \dots \\ a_{31} & a_{32} & \vdots & a_{33} \end{bmatrix}_{3\times 3} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}_{3\times 3},$$

where

$$\mathbf{A}_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}_{2\times 2}, \ \mathbf{A}_{12} = \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}_{2\times 1},$$

$$\mathbf{A}_{21} = \begin{bmatrix} a_{31} & a_{32} \end{bmatrix}_{1\times 2}, \text{ and } \mathbf{A}_{22} = \begin{bmatrix} a_{33} \end{bmatrix}_{1\times 1}.$$

A *block diagonal matrix* has square diagonal submatrices with all

off-diagonal submatrices equal to 0. For example,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}$$

is a block diagonal matrix.

## Matrix Operations

The *trace* of a square matrix $\mathbf{A}_{n\times n}$ is the sum of the diagonal elements of $\mathbf{A}$. That is, trace$(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$.

$$\text{trace}\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right) = a_{11} + a_{22} + a_{33} \qquad .$$

The *transpose* of a matrix $\mathbf{A}$, denoted $\mathbf{A}'$, changes the rows of $\mathbf{A}$ into the columns of a new matrix $\mathbf{A}'$. If $\mathbf{A}$ is an $(r \times c)$ matrix, its transpose $\mathbf{A}'$ is a $(c \times r)$ matrix.

The transpose of a column vector is a row vector, i.e.

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}_{4\times 1}' = \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix}_{1\times 4}.$$

A symmetric matrix has $\mathbf{A}' = \mathbf{A}$.

*Matrix addition* of two matrices $\mathbf{A}$ and $\mathbf{B}$ is defined only if $\mathbf{A}$ and $\mathbf{B}$ have the same number of rows and the same number of columns, *matrix addition* yields $\mathbf{A}_{r\times c} + \mathbf{B}_{r\times c} = \{a_{ij} + b_{ij}\}_{r\times c}$.

**Exercise:**

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} + \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \qquad\qquad\qquad\qquad .$$

There are several types of *matrix multiplication*. We will discuss the following types:

1. scalar multiplication,
2. matrix multiplication

- Define *scalar* multiplication of any matrix $\mathbf{A}$ by a scalar $s$ as $s\mathbf{A} = \{sa_{ij}\}$.

  **Exercise:**

$$\sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad .$$

- Two matrices $\mathbf{A}$ and $\mathbf{B}$ conform for the *matrix multiplication* $\mathbf{AB}$ if the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$. If the matrices $\mathbf{A}$ and $\mathbf{B}$ conform, *matrix multiplication* is defined as

$$\mathbf{A}_{r\times s}\mathbf{B}_{s\times t} = \left\{ \sum_{k=1}^{s} a_{ik}b_{kj} \right\} = \mathbf{C}_{r\times t}.$$

  Multiplying the $i^{th}$ row of $\mathbf{A}$ with the $j^{th}$ column of $\mathbf{B}$ yields the scalar $c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{is}b_{sj}$:

*Calculation of $c_{11}$:*

$$\text{row} \longrightarrow \begin{bmatrix} a_{11} & \rightarrow & \rightarrow & a_{1s} \\ & & & \\ & & & \end{bmatrix}_{r\times s} \begin{bmatrix} b_{11} & & \\ \downarrow & & \\ \downarrow & & \\ b_{s1} & & \end{bmatrix}_{s\times t} \downarrow \text{column}$$

So $c_{11} = a_{11}b_{11} + a_{12}b_{21} + \ldots + a_{1s}b_{s1}$.

Note that $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$. In general, $\mathbf{AB} \neq \mathbf{BA}$.

**Exercise:** Let $\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix}$.

$\mathbf{X}'\mathbf{X} =$

## Orthogonal Matrices

An *orthogonal matrix* is a **square matrix** with $\mathbf{A}' = \mathbf{A}^{-1}$. That is, a square matrix $\mathbf{A}$ is orthogonal if $\mathbf{A}'\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}'$. To establish that $\mathbf{A}$ is orthogonal, it is sufficient to show either that $\mathbf{A}'\mathbf{A} = \mathbf{I}$ or that $\mathbf{A}\mathbf{A}' = \mathbf{I}$.

- The vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal vectors* if $\mathbf{x}'\mathbf{y} = 0$.
- The vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthonormal vectors* if $\mathbf{x}$ and $\mathbf{y}$ are orthogonal vectors and are normalized: $\mathbf{x}'\mathbf{y} = 0$, $\mathbf{x}'\mathbf{x} = 1$, and $\mathbf{y}'\mathbf{y} = 1$.
- Some authors, including Muller and Fetterman, call an orthogonal matrix a column orthonormal matrix. While this is more accurate, we will retain standard terminology and refer to a matrix with $\mathbf{A}'\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}'$ as orthogonal. Thus, somewhat paradoxically, an orthogonal matrix has orthonormal columns.

## Rules of Matrix Operation

Suppose $\mathbf{A}$ and $\mathbf{B}$ conform for the operation of interest. The following laws apply to $\mathbf{A}$ and $\mathbf{B}$.

### Commutative Laws

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $a\mathbf{B} = \mathbf{B}a$

In general, $\mathbf{A}fB \neq \mathbf{BA}$.

### Distributive Laws

- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $(\mathbf{B} + \mathbf{C})\mathbf{D} = \mathbf{BD} + \mathbf{CD}$
- $a(\mathbf{B} + \mathbf{C}) = a\mathbf{B} + a\mathbf{C} = (\mathbf{B} + \mathbf{C})a$

## Associative Laws

- $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$
- $(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC})$

## Transpose Operations

- $(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$
- $(\mathbf{AB})'=\mathbf{B}'\mathbf{A}'$

## Linear Dependence and Rank

Consider the system of equations

$$\begin{aligned} 5x_1 + 2x_2 &= 2 \qquad (1) \\ 10x_1 + 4x_2 &= 4, \qquad (2) \end{aligned}$$

or

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 = \mathbf{b},$$

where

$$\mathbf{a}_1 = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \text{ and } \mathbf{b} = \begin{pmatrix} 2 \\ 4 \end{pmatrix}.$$

- This system of equations has infinite solutions given by $x_1 = \frac{2}{5} - \frac{2}{5}x_2$.
- We do not have one unique solution because the coefficients in

the second equation are a multiple of the first and thus are *linearly dependent*.

- So equation (2) does not provide any additional information about $x_1$ or $x_2$.
- Examining $\mathbf{a}_1$ and $\mathbf{a}_2$, we see that $a_{11} = \frac{5}{2}a_{12}$ and $a_{21} = \frac{5}{2}a_{22}$. Thus $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly dependent vectors.

The $n$ vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ in $\mathcal{R}^n$ are *linearly dependent* if there exist numbers $c_1, c_2, \ldots, c_n$ (not all zero) such that

$$\mathbf{a}_1 c_1 + \mathbf{a}_2 c_2 + \ldots + \mathbf{a}_n c_n = 0.$$

If this equation is true only when $c_1 = c_2 = \ldots = c_n = 0$ then the vectors are *linearly independent*.

More generally, suppose we have the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \end{bmatrix}.$$

Each column of $\mathbf{A}$ may be viewed as a vector. The *column space* of $\mathbf{A}$, denoted $C(\mathbf{A})$, is the set of all vectors that may be written as a linear combination of the columns of $\mathbf{A}$. That is, $C(\mathbf{A})$ is the set of all vectors that may be written as

$$\lambda_1 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + \lambda_2 \begin{bmatrix} 2 \\ 0 \\ 8 \end{bmatrix} + \lambda_3 \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix} = \mathbf{A} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \mathbf{A}\boldsymbol{\lambda},$$

for some vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)'$.

For example, the column space of $\mathbf{J}_2$ includes all vectors of the form

$$\lambda \mathbf{J}_2 = \begin{pmatrix} \lambda \\ \lambda \end{pmatrix},$$

including

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} -0.03 \\ -0.03 \end{pmatrix}, \text{ and } \begin{pmatrix} \pi \\ \pi \end{pmatrix}.$$

The columns of $\mathbf{A}$ are *linearly dependent* if they contain redundant information. If we can find two distinct vectors, say $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$, such that $\mathbf{A}\boldsymbol{\lambda} = \mathbf{A}\boldsymbol{\gamma} = \mathbf{x}$, then the columns of $\mathbf{A}$ are linearly dependent.

An equivalent definition, obtained by letting $\boldsymbol{\delta} = \boldsymbol{\lambda} - \boldsymbol{\gamma}$ (prove to yourself!), is that the columns of $\mathbf{A}$ are linearly dependent if there exists a vector $\boldsymbol{\delta} \neq \mathbf{0}$ such that $\mathbf{A}\boldsymbol{\delta} = \mathbf{0}$. If the columns of $\mathbf{A}$ are not linearly dependent, then they are *linearly independent*.

**Exercise:** Does the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \\ 3 & 1 & 6 \end{bmatrix}$$

have linearly independent or linearly dependent columns?

**Exercise:** Does the matrix

$$\mathbf{B} = \begin{bmatrix} 2 & 2 & 4 \\ 1 & 0 & 2 \\ 0 & 8 & 0 \\ 4 & 1 & 6 \end{bmatrix}$$

have linearly independent or linearly dependent columns?

The *rank* of a matrix $\mathbf{A}$ is the number of linearly independent columns in $\mathbf{A}$. Knowledge of the matrix rank is important in determining the existence and multiplicity of solutions to a system of linear equations.

If $\mathbf{A}$ is an $(r \times c)$ matrix with $r \geq c$, we say $\mathbf{A}$ is *full rank* if rank($\mathbf{A}$)$=c$. If rank($\mathbf{A}$)$< c$, then we say $\mathbf{A}$ is less than full rank. In linear regression, the matrix of covariates $\mathbf{X}$ must have full rank in order for our parameter estimates, $\widehat{\boldsymbol{\beta}}$, to be unique.

A square matrix that is less than full rank is called *singular*, while a full rank square matrix is called *nonsingular*. One method to find the rank of a matrix is by determining the number of linearly independent columns of a matrix.

Another method to find rank is by using elementary row operations to transform the matrix to a triangular matrix; once the matrix is in triangular form, we can determine the rank visually.

The three elementary row operations are

1. multiplying a row by a nonzero constant,
2. adding one row to another, and
3. exchanging two rows.

Other approaches to finding rank include using matrix decompositions or finding eigenvalues (the rank of a square matrix is equal to the number of nonzero eigenvalues).

## Determinants

The *determinant* is a single number summary of a square matrix that gives us information about the rank of the matrix. The determinant of a square matrix $\mathbf{A}$ is denoted $\mid \mathbf{A} \mid$ or $\det(\mathbf{A})$.

- The determinant of a diagonal or triangular matrix equals the product of the diagonal values.

  **Exercise:**

$$\left| \begin{bmatrix} 1 & 6 & 5 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix} \right| =$$

- For any $2 \times 2$ matrix,

$$\left| \begin{bmatrix} a & b \\ c & d \end{bmatrix} \right| = ad - bc.$$

- For any $3 \times 3$ matrix,

$$\begin{aligned}
\left|\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right| &= a\left|\begin{bmatrix} e & f \\ h & i \end{bmatrix}\right| - d\left|\begin{bmatrix} b & c \\ h & i \end{bmatrix}\right| + g\left|\begin{bmatrix} b & c \\ e & f \end{bmatrix}\right| \\
&= a(ei - hf) - d(bi - hc) + g(bf - ec).
\end{aligned}$$

*Useful properties of determinants*

$$
\begin{aligned}
\mid \mathbf{A}_{n\times n} \mid = 0 \quad &\Leftrightarrow \quad \text{rank}(\mathbf{A}) < n \\
&\Leftrightarrow \quad \mathbf{A} \text{ is less than full rank} \\
&\Leftrightarrow \quad \text{the } \textit{inverse} \text{ of } \mathbf{A} \text{ does not exist} \\
&\Leftrightarrow \quad \text{the columns of } \mathbf{A} \text{ are linearly dependent,}
\end{aligned}
$$

while

$$
\begin{aligned}
\mid \mathbf{A}_{n\times n} \mid \neq 0 \quad &\Leftrightarrow \quad \text{rank}(\mathbf{A}) = n \\
&\Leftrightarrow \quad \mathbf{A} \text{ is full rank} \\
&\Leftrightarrow \quad \text{the } \textit{inverse} \text{ of } \mathbf{A} \text{ exists} \\
&\Leftrightarrow \quad \text{columns of } \mathbf{A} \text{ are linearly independent.}
\end{aligned}
$$

For full rank matrices that conform, $\mid \mathbf{AB} \mid = \mid \mathbf{A} \mid\mid \mathbf{B} \mid$. In addition, $\mid \mathbf{A}' \mid = \mid \mathbf{A} \mid$.

## Positive Definite and Semidefinite Matrices

Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. $\mathbf{A}$ is *positive definite* if and only if

1. $a_{ii} > 0$ for all $i = 1, \ldots, n$
2. the determinant of every square submatrix of upper-left corner of $\mathbf{A}$ is positive. That is,

$$\left| \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \right| > 0,$$

$$\left| \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{bmatrix} \right| > 0,$$

$$\vdots$$

$$| \mathbf{A} | > 0.$$

The matrix $\mathbf{A}$ is *positive semidefinite* if we replace "$> 0$" in (1) and (2) with "$\geq 0$". A matrix is called *nonnegative definite* if it is positive definite or positive semidefinite.

Covariance matrices are nonnegative definite.

**Exercise:**

Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 & 1 \\ -1 & 4 & 0 & 2 \\ 1 & 0 & 1 & 3 \\ 1 & 2 & 3 & 2 \end{bmatrix}.$$

Is $\mathbf{A}$ positive definite, positive semidefinite, or neither?

## Inverses and Generalized Inverses

Suppose we have a system of equations like

$$(\mathbf{X}_{n\times p})'(\mathbf{X}_{n\times p})\widehat{\boldsymbol{\beta}} = (\mathbf{X}_{n\times p})'\mathbf{y}_{n\times 1},$$

where are the *normal equations* for the linear model. In order to solve these equations and obtain our estimate, $\widehat{\boldsymbol{\beta}}$, we would like to divide by $\mathbf{X}'\mathbf{X}$ in some sense. Because $\mathbf{X}'\mathbf{X}$ is a matrix, this presents us some difficulty. Thus we develop the idea of a matrix inverse.

Consider the $n \times n$ matrix $\mathbf{A}$. If $\mathbf{A}$ has full rank, then rank($\mathbf{A}$)=n, and there exists a unique matrix, $\mathbf{A}^{-1}$, called the *inverse* of $\mathbf{A}$, such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.$$

*Some properties of inverses*

1. For a scalar, $\mathbf{A}_{1\times 1} = a$, $\mathbf{A}^{-1} = \frac{1}{a}$.
2. The inverse of a diagonal matrix is the diagonal matrix of reciprocals of the diagonal elements.

3. For conforming full rank matrices, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

4. A symmetric matrix has a symmetric inverse.

5. The inverse of the transpose is the transpose of the inverse. That is, $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$. $(\mathbf{A}')^{-1}$ is also sometimes denoted $\mathbf{A}^{-T}$.

6. The determinant of the inverse is the inverse of the determinant. That is, $\mid \mathbf{A}^{-1} \mid = \frac{1}{|\mathbf{A}|}$.

7. The inverse of the $2 \times 2$ matrix $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ is given by

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

**Generalized Inverses** For certain linear models (some ANOVA models for example), we will want to obtain inverses of matrices that are not full rank. For any $\mathbf{A}_{r\times c}$, there exists a *generalized inverse*, denoted $\mathbf{A}^{-}_{c\times r}$, such that

$$\mathbf{A}_{r\times c}\mathbf{A}^{-}_{c\times r}\mathbf{A}_{r\times c} = \mathbf{A}_{r\times c}.$$

The *generalized inverse* is not unique.

The *Moore-Penrose generalized inverse (MPGI)*, $\mathbf{A}^{+}$, is a unique type of generalized inverse satisfying the following properties:

- $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$ (definition of generalized inverse)
- $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$ ($\mathbf{A}$ is a generalized inverse of $\mathbf{A}^{+}$)
- $(\mathbf{A}^{+}\mathbf{A})' = \mathbf{A}^{+}\mathbf{A}$ ($\mathbf{A}^{+}\mathbf{A}$ is symmetric)
- $(\mathbf{A}\mathbf{A}^{+})' = \mathbf{A}\mathbf{A}^{+}$ ($\mathbf{A}\mathbf{A}^{+}$ is symmetric)

*Facts about the MPGI*

1. For $\mathbf{A}_{r\times c}$ with $r \leq c$ and $\mathbf{A}$ of full row rank $r$,

$$\mathbf{A}^{+}_{c\times r} = \mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}.$$

2. For $\mathbf{A}_{r\times c}$ with $c \leq r$ and $\mathbf{A}$ of full column rank $c$,

$$\mathbf{A}^{+}_{r\times c} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'.$$

3. The Moore-Penrose generalized inverse of a less than full rank diagonal matrix is the diagonal matrix with reciprocals of the nonzero elements on the diagonal in the same locations as the nonzero elements, and zero elsewhere.

Both the Moore-Penrose generalized inverse and other generalized inverses reduce to the regular inverse when the matrix of interest is square and full rank.

## Eigenvalues, Eigenvectors, and the Spectral Decomposition

*Eigenanalysis* is defined only for square matrices. Most interest in decomposing matrices in statistics lies with symmetric matrices, for example, covariance matrices.

Suppose $\mathbf{A}$ is an $n \times n$ matrix (not necessarily symmetric). A *right eigenvector* of $\mathbf{A}$ is any nonzero $n \times 1$ vector $\mathbf{x}$ satisfying

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x},$$

where $\lambda$ is the *eigenvalue* corresponding to $\mathbf{x}$. Eigenvalues and eigenvectors are also called *characteristic values* and *characteristic vectors* (*eigen* is German for *characteristic*).

Eigenvectors are not unique (prove it?), the convention is to scale the eigenvector $\mathbf{x}$ so that $\mathbf{x}'\mathbf{x} = 1$, normalizing it to unit length.

## Finding Eigenvectors and Eigenvalues

Using the definition of eigenvectors, we can find the *characteristic equation* that is used to find eigenvalues.

$$\begin{aligned}
\mathbf{A}\mathbf{x} &= \lambda\mathbf{x} \\
\mathbf{A}\mathbf{x} - \lambda\mathbf{x} &= \mathbf{0} \\
(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} &= \mathbf{0} \\
\mid \mathbf{A} - \lambda\mathbf{I} \mid &= \mathbf{0}.
\end{aligned}$$

The last step follows because $\mathbf{x}$ is a nonzero vector.

The characteristic equation of an $(n \times n)$ matrix equals an $n^{th}$ degree polynomial in $\lambda$. An $n^{th}$ degree polynomial has $n$ roots, so a $(n \times n)$ matrix has $n$ eigenvalues. For a $(2 \times 2)$ matrix,

$$
\begin{aligned}
| \mathbf{A} - \lambda \mathbf{I} | &= \left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0 \\
\left| \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| &= \left| \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \right| = 0 \\
(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} &= 0 \\
\lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21} &= 0.
\end{aligned}
$$

Once the eigenvalues are found, eigenvectors corresponding to these eigenvalues may be found using the equation $\mathbf{Ax} = \lambda \mathbf{x}$.

**Example:**

Suppose

$$\mathbf{A} = \begin{bmatrix} 5 & -3 \\ 4 & -2 \end{bmatrix}.$$

The characteristic equation is

$$\begin{aligned} |\,\mathbf{A} - \lambda\mathbf{I}\,| &= \left| \begin{bmatrix} 5-\lambda & -3 \\ 4 & -2-\lambda \end{bmatrix} \right| \\ &= (5-\lambda)(-2-\lambda) - (-3)(4) \\ &= \lambda^2 - 3\lambda + 2 \\ &= (\lambda - 2)(\lambda - 1) = 0, \end{aligned}$$

so the eigenvalues of $\mathbf{A}$ are $2$ and $1$.

We find eigenvectors corresponding to these eigenvalues by solving $\mathbf{Ax} = \lambda\mathbf{x}$ for $\lambda = 1$ and $\lambda = 2$.

For $\lambda = 1$, we have

$$\begin{bmatrix} 5 & -3 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

which leads to the system of equations

$$5x_1 - 3x_2 = x_1$$
$$4x_1 - 2x_2 = x_2.$$

Rearranging both equations, we see $x_1 = \frac{3}{4}x_2$. So one eigenvector corresponding to the eigenvalue 1 is $(3, 4)'$.

To normalize this vector, we take

$$\begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 9 + 16 = 25$$

and divide by the square root of this number. So the normalized eigenvector corresponding to the eigenvalue 1 is $(\frac{3}{5}, \frac{4}{5})'$.

Similarly, we can show the normalized eigenvector corresponding to $\lambda = 2$ is $\frac{1}{\sqrt{2}}(1, 1)'$.

## Some Properties of Eigenvalues and Eigenvectors

- For $\mathbf{A}_{n\times n}$, the number of distinct eigenvalues ranges from 1 to $n$.
- The trace of a matrix is the sum of the eigenvalues. That is, trace$(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$.
- The determinant of a matrix is the product of its eigenvalues. That is, $\mid \mathbf{A} \mid = \prod_{i=1}^{n} \lambda_i$.
- $\mathbf{A}$ full rank $\Leftrightarrow$ $\mathbf{A}$ has no zero eigenvalues
- $\mid \mathbf{A} \mid = 0 \Leftrightarrow$ at least one eigenvalue is zero $\Leftrightarrow$ $\mathbf{A}$ is not full rank
- The number of nonzero eigenvalues of $\mathbf{A}$ is rank$(\mathbf{A})$.
- Small eigenvalues imply that there are near-linear dependencies in the columns of a matrix $\mathbf{A}$.
- $\mathbf{A}$ is positive definite if min$(\lambda_i) > 0$
- $\mathbf{A}$ is positive semidefinite if min$(\lambda_i) \geq 0$

## Spectral Decomposition

It is sometimes easier to deal with matrices if we write them as a product of more simple matrices that have special structures. Matrix decomposition allows us to write matrices as a product of simpler matrices. We will see one such decomposition, the *spectral decomposition*, later in the course.

The *spectral decomposition* allows us to write any symmetric matrix in terms of an orthogonal matrix and a diagonal matrix of eigenvalues.

*Spectral Theorem:* Suppose $\mathbf{A}$ is an $(n \times n)$ symmetric matrix. Then there exists an orthogonal (column orthonormal) matrix $\mathbf{V}$ such that

$$\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}',$$

where $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is an $(n \times n)$ diagonal matrix of the ordered eigenvalues of $\mathbf{A}$ so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. $\mathbf{V}$ is the orthogonal matrix of eigenvectors corresponding to the eigenvalues of $\mathbf{A}$. The eigenvalues and eigenvectors must be in the same order.

**Example:** Verify that the spectral decomposition of

$$\mathbf{A} = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$

is given by $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$, where

$$\mathbf{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \text{ and } \boldsymbol{\Lambda} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Once a matrix is in spectral decomposition form, it is easy to obtain the inverse.

- For $\mathbf{A}$ full rank, we have $\mathbf{A}^{-1} = \mathbf{V}(\mathbf{\Lambda})^{-1}\mathbf{V}'$.
- For $\mathbf{A}$ less than full rank, we have $\mathbf{A}^{+} = \mathbf{V}(\mathbf{\Lambda})^{+}\mathbf{V}'$.

Thus computing the reciprocals of the eigenvalues allows us to find the inverse or Moore-Penrose generalized inverse of a symmetric matrix in this form.

We will see the spectral decomposition again in Chapter 8.

## Singular Value Decomposition (SVD)

The *singular value decomposition* gives us a more accurate way to find the inverse of an ill-conditioned matrix. Both SAS and S-plus use the SVD when fitting linear models. The SVD is valid for *any* matrix, while the spectral decomposition is valid only for symmetric matrices.

For *any* matrix $\mathbf{A}_{m\times n}$ with $m \geq n$, there exist orthogonal matrices $U_{m\times m}$ and $V_{n\times n}$ along with an $(m \times n)$ matrix $\mathbf{S}$ such that

$$\mathbf{A}_{m\times n} = \mathbf{U}_{m\times m}\mathbf{S}_{m\times n}\mathbf{V}'_{n\times n}.$$

We define

$$\mathbf{S}_{m\times n} = \begin{bmatrix} \text{Diag}(\mathbf{s}) \\ \mathbf{0}_{(\mathbf{m}-\mathbf{n})\times\mathbf{n}} \end{bmatrix},$$

where the vector $\mathbf{s}_{n\times 1}$ contains the $n$ *singular values* of $\mathbf{A}$. The rank $r$ of $\mathbf{A}$ is the number of nonzero singular values of $\mathbf{A}$. Singular values are computed as the positive square roots of the eigenvalues of $\mathbf{A}'\mathbf{A}$.

*Facts about the SVD*

- $\mathbf{U}'\mathbf{U} = I_m \qquad \mathbf{V}'\mathbf{V} = I_n$
- $\mathbf{A}'\mathbf{A} = \mathbf{V}\text{Diag}(\mathbf{s}^2)\mathbf{V}'$ (note: $\text{Diag}(\mathbf{s}^2)$ here means $\text{Diag}(\{s_i^2\})$)
- $\mathbf{A}\mathbf{A}' = \mathbf{U}\text{Diag}(\mathbf{s}^2, \mathbf{0_{m-n}})\mathbf{U}'$
- The Moore-Penrose generalized inverse of $\mathbf{A}$ may be written

$$\begin{aligned} \mathbf{A}^+ &= \mathbf{V}\mathbf{S}^+\mathbf{U}' \\ &= \mathbf{V}\left[\text{Diag}(s)^+ \quad \mathbf{0_{n\times(m-n)}}\right]\mathbf{U}', \end{aligned}$$

  where $\text{Diag}(s)^+ = \text{Diag}(\{s_1^{-1}, s_2^{-1}, \ldots, s_r^{-1}, 0, \ldots, 0\})$.

- SVD is also widely used in statistics where it is related to principal component analysis, and in signal processing and pattern recognition.

## Random Vectors and Matrices

Suppose

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

is a vector of random variables with $E(Y_i) = \mu_i$, $\text{Var}(Y_i) = \sigma_{ii}$, and $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$.

The expectation of the random vector $\mathbf{Y}$ is defined

$$E(\mathbf{Y}) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}.$$

Suppose $\mathbf{Z}$ is an $(n \times p)$ matrix of random variables. Then

$$E(\mathbf{Z}) = \begin{pmatrix} E(Z_{11}) & \dots & E(Z_{1p}) \\ \vdots & \dots & \vdots \\ E(Z_{n1}) & \dots & E(Z_{np}) \end{pmatrix}.$$

Thus the expectation of a random matrix is the matrix of the expectations.

For $\mathbf{Y}$ an $(n \times 1)$ random vector, the *covariance matrix* of $\mathbf{Y}$ is

$$\begin{aligned}
\text{Cov}(\mathbf{Y}) &= E\left[(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})'\right] \\
&= \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \ldots & \ldots & \sigma_{nn} \end{pmatrix},
\end{aligned}$$

where $\sigma_{ij} = E[(Y_i - \mu_i)(Y_j - \mu_j)']$, $i, j = 1, \ldots, n$.

Suppose $\boldsymbol{\mu} = E(\mathbf{Y})$ and covariance matrix $\boldsymbol{\Sigma} = \text{Cov}(\mathbf{Y})$. In addition, suppose $\mathbf{A}_{r\times n}$ is a matrix of constants and $\mathbf{b}_{r\times 1}$ is a vector of constants. Then

$$\begin{aligned}
E(\mathbf{AY}+\mathbf{b}) &= \mathbf{A}E(\mathbf{Y}) + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\
\text{Cov}(\mathbf{AY}+\mathbf{b}) &= \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.
\end{aligned}$$

Let $\mathbf{W}_{r\times 1}$ be a random vector with $E(\mathbf{W}) = \boldsymbol{\gamma}$. Then

$$\text{Cov}(\mathbf{W}, \mathbf{Y}) = E\left[(\mathbf{W} - \boldsymbol{\gamma})(\mathbf{Y} - \boldsymbol{\mu})'\right],$$

where $\text{Cov}(\mathbf{W}, \mathbf{Y})$ is an $(r \times n)$ matrix of covariances with $ij^{th}$ element equal to $\text{Cov}(W_i, Y_j)$.

## Important Distributions for Linear Models

The theory of linear models involves many statistical distributions, including the following distributions:

1. Gaussian (Normal) Distribution
2. Multivariate Normal Distribution
3. Chi-Squared Distribution
4. Non-Central Chi-Squared Distribution
5. t Distribution
6. Non-Central t Distribution
7. F Distribution
8. Non-Central F Distribution.

We will now discuss many of these distributions in detail.

## Gaussian (Normal) Distribution

A random variable $X$ has a Gaussian (normal) distribution with mean $\mu$ and variance $\sigma^2$, written $X \sim N(\mu, \sigma^2)$, if $X$ has density

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

When $\mu = 0$ and $\sigma^2 = 1$, then we say $X \sim N(0, 1)$ has a *standard normal distribution*.

If $X \sim N(\mu, \sigma^2)$, then $\frac{x-\mu}{\sigma} \sim N(0, 1)$. For the standard normal distribution, 95% of the probability mass falls between -1.96 and 1.96 (roughly 2 standard deviations of the mean). Much of hypothesis testing is based on this fact.

The normal distribution is symmetric about its mean.

$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$
$\varphi_{\mu,\sigma^2}(x)$
$x$
1.0
0.8
0.6
0.4
0.2
0.0
−5
−4
−3
−2
−1
0
1
2
3
4
5

## Multivariate Normal Distribution

Suppose $X = (X_1, \ldots, X_n)'$. Then $X$ has an $n$ dimensional multivariate normal distribution with mean $\mu$ and covariance matrix $\Sigma$ if $X$ has density

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \mid \Sigma \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}.$$

We write this $X \sim N_n(\mu, \Sigma)$.

This definition requires $\Sigma$ to be positive definite.

0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
0
10
20
30
40
50
60
70
80
90
100
X
0
10
20
30
40
50
60
70
80
90
Y

*Facts about the multivariate normal distribution*

1. A linear transformation of a multivariate normal distribution yields another multivariate normal distribution. Suppose $X \sim N_n(\mu, \Sigma)$. For $A_{r\times n}$ a matrix of constants and $b_{r\times 1}$ a vector of constants, then $Y = AX + b$ has the multivariate normal distribution given by $Y \sim N_r(A\mu + b, A\Sigma A')$.

2. A linear combination of independent multivariate normal distributions is a multivariate normal distribution. Suppose $X_1, \ldots, X_k$ are independent with $X_i \sim N_n(\mu_i, \Sigma_i)$, $i = 1, \ldots, k$. Suppose $a_1, \ldots, a_k$ are scalars and define

$$Y = a_1 X_1 + \ldots + a_k X_k.$$

   Then $Y \sim N(\mu^*, \Sigma^*)$, where $\mu^* = \sum_{i=1}^k a_i \mu_i$ and $\Sigma^* = \sum_{i=1}^k a_i^2 \Sigma_i$.

3. Marginal distributions of a multivariate normal distribution are multivariate normal distributions. Suppose $X \sim N_n(\mu, \Sigma)$.

Partition $X$ into $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ where $X_1$ is $r \times 1$ and $X_2$ is $(n-r) \times 1$. Partition $\mu$ as $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ where $\mu_1$ is $r \times 1$ and $\mu_2$ is $(n-r) \times 1$. Similarly, partition $\Sigma$ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where $\Sigma_{11}$ is $r \times r$, $\Sigma_{12}$ is $r \times (n-r)$, $\Sigma_{21} = \Sigma_{12}'$ is $(n-r) \times r$, and $\Sigma_{22}$ is $(n-r) \times (n-r)$. Then the marginal distribution of $X_1$ is given by $X_1 \sim N_r(\mu_1, \Sigma_{11})$, and the marginal distribution of $X_2$ is given by $X_2 \sim N_{(n-r)}(\mu_2, \Sigma_{22})$.

4. Conditional distributions of multivariate normal distributions are multivariate normal distributions. Suppose that $X \sim N_n(\mu, \Sigma)$.

Using the same partition as above, then we have

$$\begin{aligned} X_1 \mid X_2 &= x_2 \\ &\sim N_r(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma^*), \end{aligned}$$

where $\Sigma^* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

## Chi-Squared Distribution

A random variable $X$ has a central chi-squared distribution with $n$ *degrees of freedom*, written $X \sim \chi^2(n)$, if the density of $X$ is given by

$$f(x) = \left(\frac{1}{\Gamma(\frac{n}{2})}\right)\left(\frac{1}{2}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \exp\left\{-\frac{x}{2}\right\},$$

where $\Gamma(a)$ is the complete gamma function, given by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

The chi-squared distribution is asymmetric and restricted to positive numbers. Its degrees of freedom determine the mean and variance of the distribution.

1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
k=1
k=2
k=3
k=4
k=5

The chi-squared distribution is related to the normal distribution. If the random variable $Z \sim N(0,1)$, then $Z^2 \sim \chi^2(1)$. In addition, if $Z_1, Z_2, \ldots, Z_n$ are independent, identically distributed $N(0,1)$ random variables, then $W = \sum_{i=1}^{n} Z_i^2$ has a chi-squared distribution with $n$ degrees of freedom; that is, $W \sim \chi^2(n)$.

The mean of a $\chi^2(n)$ distribution is $n$, and its variance is $2n$.

If $Z \sim N(\mu, 1)$, then $Z^2$ follows a non-central chi-squared distribution with 1 degrees of freedom and non-centrality parameter $\mu^2$.

The chi-squared distribution is used widely in the analysis of categorical data.

## Student's t Distribution

Suppose $X \sim N(0,1)$, $Y \sim \chi^2(n)$, with $X$ and $Y$ independent. The random variable

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

has a $t$ distribution with $n$ degrees of freedom. We write this as $T \sim t(n)$. Like the standard normal distribution, the t distribution is symmetric about 0.

0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
k=1
k=2
k=5
k=10
k=infini

If $X \sim N(\mu, 1)$, then $T$ has a non-central t-distribution with $n$ degrees of freedom and non-centrality parameter $\mu$.

The *degrees of freedom* $n$ determines the amount of variability in the distribution. As the number of degrees of freedom increases, the variability of the $t$ distribution decreases. In fact, as the number of degrees of freedom gets large, the $t$ distribution approximates the standard normal distribution. With smaller degrees of freedom, the t distribution resembles a normal distribution with fatter tails.

A $t(1)$ distribution, which has 1 degree of freedom, is not well-behaved and is called a *Cauchy distribution*.

## Fisher's F Distribution

Suppose $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, and $X_1$ and $X_2$ are independent. The random variable

$$F = \frac{\left(\frac{X_1}{n_1}\right)}{\left(\frac{X_2}{n_2}\right)}$$

has a central $F$ distribution with $(n_1, n_2)$ degrees of freedom. We write this $F \sim F(n_1, n_2)$.

We call $n_1$ the *numerator* degrees of freedom and $n_2$ the *denominator* degrees of freedom. The central $F$ distribution is used in hypothesis tests of nested linear models. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$. The $F$ distribution is asymmetric and restricted to positive numbers.

If $X_1 \sim \chi^2(n_1; \mu^2)$, then $F$ follows a non-central F distribution.

2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

## Maximum Likelihood Estimation

*Maximum likelihood estimates (MLE's)* have excellent large-sample properties and are applicable in a wide variety of situations. Examples of maximum likelihood estimates include the following:

- The sample average $\bar{X}$ of a group of independent and identically normally distributed observations $X_1, \ldots, X_n$ is a MLE.
- Parameter estimates in a linear regression model fit to normally distributed data are maximum likelihood estimates.
- Parameter estimates in a logistic regression model are maximum likelihood estimates.
- The estimate $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$ of the variance of a group of independent and identically normally distributed observations is *not* a MLE. (The MLE of the variance is $\left(\frac{n-1}{n}\right) s^2$.)

## Finding Maximum Likelihood Estimates

Let $L(\mathbf{Y} \mid \boldsymbol{\theta})$ denote the *likelihood function* for $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ from some population described by the parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

The maximum likelihood estimate of $\boldsymbol{\theta}$ is given by the estimator $\widehat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)$ for which

$$L(\mathbf{Y} \mid \widehat{\boldsymbol{\theta}}) > L(\mathbf{Y} \mid \boldsymbol{\theta}^*),$$

where $\boldsymbol{\theta}^*$ is any other estimate of $\boldsymbol{\theta}$.

Thus the maximum likelihood estimate is the most "probable" or "likely" for the data.

Maximizing the likelihood $L(\mathbf{Y} \mid \boldsymbol{\theta})$ is equivalent to maximizing the natural logarithm $\ln(L(\mathbf{Y} \mid \boldsymbol{\theta})) = \ell(\mathbf{Y} \mid \boldsymbol{\theta})$, called the log-likelihood.

The maximum likelihood estimates are typically found as the solutions of the $p$ equations obtained by setting the $p$ partial derivatives of $\ell(\mathbf{Y} \mid \boldsymbol{\theta})$ with respect to each $\boldsymbol{\theta}_j$, $j = 1, \ldots, p$, equal to zero.

Why do we solve the derivatives for zero? The derivative gives us the slope of the likelihood (or log-likelihood), and when the slope is zero, we know that we are at either a local minimum or local maximum. (The second derivative is negative for a maximum and positive for a minimum.)

When closed form expressions for maximum likelihood estimates do not exist, computer algorithms may be used to solve for the estimates.

**Example:**

Let $Y_i$, $i = 1, \ldots, n$ be i.i.d.normal random variables with mean $\mu$ and variance $\sigma^2$, so $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$.

The density of $Y_i$ is given by

$$f(Y_i \mid \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \mu)^2 \right\}.$$

Find the maximum likelihood estimates of $\mu$ and $\sigma^2$.

## Hypothesis Testing and Interval Estimation with MLE's

It can be shown (based on large-sample properties of MLE's) that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_1)}}$$

is approximately $N(0,1)$ when the sample size is large.

A test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ can be based on the $Z$ statistic $\frac{\hat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_1)}}$, which has approximately a $N(0,1)$ distribution under $H_0$. This test is called a *Wald test*.

By a similar argument, an approximate $100(1-\alpha)\%$ large-sample confidence interval for $\beta_1$ takes the form

$$\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathsf{Var}}(\hat{\beta}_1)},$$

where $Pr(Z > Z_{1-\frac{\alpha}{2}}) = \frac{\alpha}{2}$ when $Z \sim N(0,1)$.

*NOTE:* Testing $\sigma^2$ is more complicated, and the Wald test for the hypothesis $\sigma^2 = 0$ is not recommended because the value $\sigma^2 = 0$ is on the boundary of the parameter space for $\sigma^2$.

## Next: Simple Linear Regression

*Reading Assignment:*

- Weisberg Chapter 2: Simple Linear Regression

# Introduction to R

Also, see cran.**r**-project.org/doc/manuals/**R**-intro.pdf

A schematic view of R (from E. Paridis, R for Beginners, 2002)

# Why R?

- free!
- runs on a variety of platforms including Windows, Unix and MacOS.
- provides an unparalleled platform for programming new statistical methods in an easy and straightforward manner.
- contains advanced statistical routines not yet available in other packages.
- has state-of-the-art graphics capabilities.

# R Tutorials

1. **P. Kuhnert & B. Venables, An Introduction to R: Software for Statistical Modeling & Computing**
2. **J.H. Maindonald, Using R for Data Analysis and Graphics**
3. **B. Muenchen, R for SAS and SPSS Users**
4. **W.J. Owen, The R Guide**
5. **D. Rossiter, Introduction to the R Project for Statistical Computing for Use at the ITC**
6. **W.N. Venebles & D. M. Smith, An Introduction to R**

- Variables, data, functions, results, etc, are stored in the active memory of the computer in the form of *objects* which have a *name*. The user can do actions on these objects with *operators* (arithmetic, logical, and comparison) and *functions* (which are themselves objects)

- Let's illustrate by starting with the command line in a new session. You'll see ">" starting each line
- To quit R, use >q()

# R Overview

- Most functionality is provided through built-in and user-created functions and all data objects are kept in memory during an interactive session
- Basic functions are available by default. Other functions are contained in packages that can be attached to a current session as needed

Show current *working directory*, where data objects will be stored

```
> getwd()
>setwd("C:\\Users\\feizou\\Desktop\\Teachin
g\\Bios663-2019")
```

Sets working directory. It's a good idea to manage different projects separately into different working directories.

```
> # comments follow the '#' sign
> a=10 # R ignores everything after '#'
```

```
> n
Error: Object "n" not found
> n=10
> n
[1] 10
> n<-15
> n
[1] 15
> n=15; a=20
```

I typed “n” and hit return

Doesn’t exist yet

Now I assign (using `=` or `=`) the value 10 to object n

Now it exists

Alternate way of assigning.

Multiple commands on same line using semicolon

If the object already exists, its previous value is erased (the modification affects only the objects in the active memory, not the data on the disk).
Note: R is a case sensitive language. N and N are two different objects

```
> n==20
[1] FALSE
> n==15
[1] TRUE
> n<10
[1] FALSE
> n>10
[1] TRUE
> n==n
[1] TRUE
> a<-15
> a< -15
[1] FALSE
> n=a
>n==a
[1] TRUE
```

An assignment.

A query.

An assignment.

A query.

*Functions* need an argument, provided in parentheses

```
> a=10

> sqrt(a)

> [1] 3.162278
```

Object names cannot start with digits, but they can have digits within them. For long names, it's a good idea to use separators such as "." or "_"

Names that are already in use by R (usually as functions) are a bad idea. Single letter names could be a problem, like "t" (used for the transpose function, etc.). But it's easy to check.

# Removing/deleting objects from the current working directory

```
> objects()
> ls()
> rm(n)
> objects()
```

Similar

Object is removed

```
> d
Error: Object "d" not found
> t
function (x)
UseMethod("t")
<environment: namespace:base>
```

# Variable data **types** – character, numeric, logical

```
> a="scooby doo"
> a
[1] "scooby doo"
> is.character(a)
[1] TRUE
> typeof(a)
[1] "character"
```

CHARACTER

functions that start with “is” often can answer a question about an object

```
> a=10
> is.numeric(a)
[1] TRUE
> is.character(a)
[1] FALSE
>
> a=10/sqrt(2)
> a
[1] 7.071068
> is.numeric(a)
[1] TRUE
> a^100
[1] 8.881784e+84
> a^1000
[1] Inf
> a^(-1000)
[1] 0
```

NUMERIC

Too big

Too tiny

```
> a=T
> a
[1] TRUE
> a=TRUE
> a
[1] TRUE
> a+1
[1] 2
```

LOGICAL

Treats TRUE as 1, FALSE as zero when numeric operations performed

# R as a calculator

```
> a=2
> b=10
> c=3
>
> (-b+sqrt(b^2-4*a*c))/(2*a)
[1] -0.3205505
> (-b-sqrt(b^2-4*a*c))/(2*a)
[1] -4.679449
```

# R object types

- **Vector:**
  - a one-dimensional array of arbitrary length. Subsets of the vector may be referenced. All elements of the vector must be of the same data type--numerical, character, etc.
- **Matrix:**
  - a two-dimensional array with an arbitrary number of rows and columns. Subsets of the matrix may be referenced, and individual rows and columns of the matrix may be handled as vectors. Again all elements of the matrix must be of the same data type.
- **Array:**
  - as a matrix, but of arbitrary dimension.

Source: http://www.ma.hw.ac.uk/~stan/R/Rnotes.pdf

So far we have assigned a single element to each object. This is a special case of a vector.

The “c” function concatenates things into a single vector

```
> a=c(1,2,3,4)
> a
[1] 1 2 3 4
> a=(1:4)
> a
[1] 1 2 3 4
> typeof(a)
[1] "integer"
> is.numeric(a)
[1] TRUE
> is.vector(a)
[1] TRUE
>
```

Special case of numeric type

So far we have assigned a single element to each object. This is a special case of a vector.

```
> a
[1] 1 2 3 4
> a=a+1
> a
[1] 2 3 4 5
> a=c(a, "scooby doo")
> a
[1] "2"          "3"          "4"
"5"          "scooby doo"
> typeof(a)
[1] "character"
>
```

What does this do?

We say that a was “coerced” into the character data type

# Simple plots

```
> a=2
> b=10
> c=3
> x=c(-50:50)/10
>
> y=a*x^2+b*x+c
> plot(x,y)
> abline(0,0)
```

Vector from what value to what value?

Another vector

Simple scatterplot – two vectors must have elements in the same order

Draws a line on a plot with intercept 0 and slope 0

Matrix outer product

```
> a=c(1:11)
> b=c(-5:5)
> d=outer(a,b,FUN="*")
> d
      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]
 [1,]   -5   -4   -3   -2   -1    0    1    2    3     4     5
 [2,]  -10   -8   -6   -4   -2    0    2    4    6     8    10
 [3,]  -15  -12   -9   -6   -3    0    3    6    9    12    15
 [4,]  -20  -16  -12   -8   -4    0    4    8   12    16    20
 [5,]  -25  -20  -15  -10   -5    0    5   10   15    20    25
 [6,]  -30  -24  -18  -12   -6    0    6   12   18    24    30
 [7,]  -35  -28  -21  -14   -7    0    7   14   21    28    35
 [8,]  -40  -32  -24  -16   -8    0    8   16   24    32    40
 [9,]  -45  -36  -27  -18   -9    0    9   18   27    36    45
[10,]  -50  -40  -30  -20  -10    0   10   20   30    40    50
[11,]  -55  -44  -33  -22  -11    0   11   22   33    44    55
```

A single element

```
> d[3,2]
[1] -12
```

Both are vectors

A column

```
> d[,2]
 [1]  -4  -8 -12 -16 -20 -24 -28 -32 -36 -40
-44
```

A row

```
> d[3,]
 [1] -15 -12  -9  -6  -3   0   3   6   9  12
15
```

```
> image(d)
```

Image of a matrix, color corresponds to matrix elements

```
> smaller.d=d[2:3,4:8]
> smaller.d
     [,1] [,2] [,3] [,4] [,5]
[1,]   -4   -2    0    2    4
[2,]   -6   -3    0    3    6
> smaller.d^2
     [,1] [,2] [,3] [,4] [,5]
[1,]   16    4    0    4   16
[2,]   36    9    0    9   36
> temp.matrix=matrix(c(1:10),5,2)
> temp.matrix
     [,1] [,2]
[1,]    1    6
[2,]    2    7
[3,]    3    8
[4,]    4    9
[5,]    5   10
```

Subset of matrix is a matrix (unless one of the dimensions is 1, in which case it's a vector)

Many operations operate element-wise

rows

columns

Q: what is `temp.matrix[2,]`?

# The apply function

```
> temp.matrix
     [,1] [,2]
[1,]    1    6
[2,]    2    7
[3,]    3    8
[4,]    4    9
[5,]    5   10
> apply(temp.matrix,1,mean)
[1] 3.5 4.5 5.5 6.5 7.5
> rowMeans(temp.matrix)
[1] 3.5 4.5 5.5 6.5 7.5
> rowSums(temp.matrix)
[1]  7  9 11 13 15
> apply(temp.matrix,1,var)
[1] 12.5 12.5 12.5 12.5 12.5
> apply(temp.matrix,2,var)
[1] 2.5 2.5
```

Faster way to do the same thing

No analogous function rowVars to speed things up

# Matrix

```
A <- matrix(c(1,2,4,1), ncol=2)
B<- matrix(c(1,4,5,6), ncol=2)
A*B
A%*%B
t(A)
det(A)
solve(A)
eigen(A)
```

```
> temp.array=array(1:24,c(2,3,4))
> temp.array
, , 1

     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

, , 2

     [,1] [,2] [,3]
[1,]    7    9   11
[2,]    8   10   12

, , 3

     [,1] [,2] [,3]
[1,]   13   15   17
[2,]   14   16   18

, , 4

     [,1] [,2] [,3]
[1,]   19   21   23
[2,]   20   22   24
```

Q: what is `temp.array[2,2,2:3]`?

# The apply function on arrays

```
> temp.array=array(1:24,c(2,3,4))
> apply(temp.array,1,mean)
[1] 12 13
> apply(temp.array,2,mean)
[1] 10.5 12.5 14.5
> apply(temp.array,c(1,2),mean)
     [,1] [,2] [,3]
[1,]   10   12   14
[2,]   11   13   15
> apply(temp.array,c(1,3),mean)
     [,1] [,2] [,3] [,4]
[1,]    3    9   15   21
[2,]    4   10   16   22
>
```

# Factors and tapply

Apply a function to
"ragged" arrays

```
>a=c("NC","NC","NY","NC","GA","GA","NY","FL","GA","GA",
"SC","SC","NY","NY")
> scores=c(0.33,0.00,-1.75,0.74,-0.30,-0.36,-
0.57,0.83,0.75,-1.42,-0.83,-0.19,-0.10,1.68)

> tapply(scores,a,FUN=mean)
        FL          GA          NC          NY          SC
 0.8300000 -0.3325000  0.3566667 -0.1850000 -0.5100000
> tapply(scores,a,FUN=sd)
       FL        GA        NC        NY        SC
       NA 0.8862421 0.3707200 1.4239499 0.4525483
>
```

Missing value

# Simple histograms and boxplots

```
> hist(scores)
> boxplot(scores)
> boxplot(scores~a.f)


> plot(scores,a.f)
> plot(a.f,scores)
```

# R object types, cont.

- **Data frame:**
  - a set of **data** organized similarly to a matrix. However, each column of the data frame may contain its own type of data. Columns typically correspond to variables in a statistical study, while rows correspond to observations of these variables. A data frame may be handled similarly to a matrix, and individual columns of the data frame may be handled as vectors.
- **List:**

Powerful feature of R

  - an arbitrary collection of other **R** objects (which may include other lists).

Source: http://www.ma.hw.ac.uk/~stan/R/Rnotes.pdf

# R object types, cont.

- **function**:

  - Doesn’t have a data type.  This is an object that takes other objects as arguments and does something with them.

# Random variables and distributions

| Distribution | R name | additional arguments |
|---|---|---|
| beta | `beta` | `shape1, shape2, ncp` |
| binomial | `binom` | `size, prob` |
| Cauchy | `cauchy` | `location, scale` |
| chi-squared | `chisq` | `df, ncp` |
| exponential | `exp` | `rate` |
| F | `f` | `df1, df2, ncp` |
| gamma | `gamma` | `shape, scale` |
| geometric | `geom` | `prob` |
| hypergeometric | `hyper` | `m, n, k` |
| log-normal | `lnorm` | `meanlog, sdlog` |
| logistic | `logis` | `location, scale` |
| negative binomial | `nbinom` | `size, prob` |
| normal | `norm` | `mean, sd` |
| Poisson | `pois` | `lambda` |
| Student's t | `t` | `df, ncp` |
| uniform | `unif` | `min, max` |
| Weibull | `weibull` | `shape, scale` |
| Wilcoxon | `wilcox` | `m, n` |

# Random variables and distributions, cont.

Random variable  Quantile  Density

## Example: rnorm, qnorm, dnorm, etc.

```
> plot(rnorm(100,0,1),rnorm(100,0,1))
> x=rnorm(100,0,1)
> y=x+rnorm(100,0,1)
> x=c(-50:50)/10
> y=dnorm(x,mean=0,sd=1)
> plot(x,y)
> y=pnorm(x,mean=0,sd=1)
> plot(x,y)
```

# Searching for information/functions

```
> help(regression)
No documentation for 'regression' in
specified packages and libraries:
you could try 'help.search("regression")'
>
>
> help("lsfit")
> help.search("regression")
```

Now we're getting somewhere (see popup)

Searches for anything with `regression' in it

## R hints from Karl Broman's website, with additions

* Use the "recording" feature of the graphics window.
After making a plot, select (from the menu bar) History:Recording. After making further plots, you may use the Page Up and Page Dn keys to go back and forth through the previous plots you've made.

* Save your analysis commands in a text file.
Use a text editor (such as Notepad or an enhanced version such as EditPad) to edit a file containing your commands. Use the Windows keyboard shortcut Alt:Tab to quickly switch between R and your text editor.

Alternatively, use File->New Script for an R Editor.

Note that commands you type in R are saved (temporarily) in the file .Rhistory (in the same directory as your .RData file). You may wish to peruse (or copy and paste from) this file and keep a permanent record your analysis commands.

* To delete all objects in your workspace, type (from the R prompt)

```
rm(list=ls())
```

This is useful in debugging to ensure that your code does not rely on existing values for data objects, so that you can mimic what a naïve user would encounter by using your code. However, it's also very dangerous, as it removes all objects from your workspace.

## Additional Resources

http://cran.r-project.org/doc/manuals/R-intro.html
http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf
Quick reference card: http://people.musc.edu/~slateeh/Rintro/RQuickReference.pdf

```
# VERY BASIC THINGS
x = 1:10
x
y = 2
z = x + y
logi = x < 4
logi
mode(x)
mode(y)
mode(logi)
as.numeric(logi)
mat = matrix(z, ncol = 2)
mat
mat2 = matrix(z, ncol = 2, byrow = TRUE)
dim(mat)
mat[1,2]
mat[,1]
```

```
# VERY BASIC THINGS, cont.
mat[2,]
colnames(mat) = paste("col", 1:2)
colnames(mat)
mat
mat[,"col 2"]
t(mat)
arraydat = array(1:27, dim = c(3,3,3))
arraydat
mylist = list(x = x, y = y, z = z, mat = mat, arrayd
= arraydat)
mylist[["arrayd"]]
mylist$arrayd
mylist$x
mylist[["y"]]
mydata = data.frame(x = x, y = y, logi = logi)
names(mydata)
```

transpose

Creating a list

Double brackets used to indicate list “elements”

Name

Sub-object assigned to that name

```
# CREATING AND MANIPULATING DATA
a = c(1,2,3)
a^2
a + 23
b = 4:6
d = seq(8, 12, by = 2)
e = c(a,b)
f = rep(a, 3)
g = rep(a, b)
m1 = cbind(a, b)
m2 = rbind(a, b)
matrix(e, ncol = 2)
matrix(e, ncol = 2, byrow = TRUE)
h = list(x = a, y = d)
apply(m1, 2, sum)  ### sum the columns
lapply(f, sum)### sum the elements in each list item
lapply(h, sum)
colors = c("blue", "green", "red")
paste("My favorite color is", colors)
```

A sequence of values

Repeating values

Column and row “bind” statements make matrices

```
# CREATING AND MANIPULATING DATA, CONT.
mydata = data.frame(aa = a, bb = b, dd = d)
mydata$aa
mydata["aa"]
mydata$zz = mydata$aa + mydata$bb
mydata$xx    ### NULL, doesn't exist
mydata$ff = g    ### wrong because g is length 15
mydata    ### error given above, nothing has been changed
a < 2
a < c(0,4,2)
a < 2 & a > 1
z = c(17, 2, 4, 23, 40, 4)
z + c(0,1,2)
sum(z)
prod(z)
max(z)
z[z > 4]
z[z==max(z)]
which.max(z)
unique(z)
```

Logical "and" – see ?Logic

Subsetting by logicals

Returns index of max

Unique set of elements within object

```
#EXAMPLES OF PLOTS

x=rnorm(100)
y=rnorm(100)+x
plot(x,y)

x=rnorm(1e5,0,2)
y=rnorm(1e5,1,3)+x
plot(x,y,pch=".")

x=seq(1,20,by=.5)
y=2^x-x^2
plot(x,y,type="b")
plot(x,log(y))
plot(log(x),log(y))
z=log(y)
print(is.na(z))
grep(T,is.na(z))
```

Change print character

is.na returns logical vector for missing or not

return indices for which the statement is true

**#EXAMPLES OF PLOTS, CONT.**

```
z = lm(dist ~ speed, data = cars)
cars
plot(cars)
plot(cars$speed,cars$dist)
abline(z) # equivalent to abline(reg = z) or
abline(coef = coef(z))
cars.loess=loess(dist ~ speed, data = cars)
lines(cars$speed,cars.loess$fitted,lwd=3)
# what is the effect of dropping the maximum observation?
which.biggest=which.max(cars$dist)
z.new=lm(dist[-which.biggest] ~ speed[-which.biggest], data
= cars)
abline(coef = coef(z.new),col="green")
cars.loess.new=loess(dist[-which.biggest] ~ speed[-
which.biggest], data = cars)
lines(cars$speed[-
which.biggest],cars.loess.new$fitted,col="green",lwd=3)
```

```
#PLOTS, CONT.
plot(cars)
plot(cars$speed,cars$dist)
abline(z)
which.identified=identify(cars,n=1)
z.new=lm(dist[-which.identified] ~ speed[-
which.identified], data = cars)
abline(coef = coef(z.new),lty=2)
points(cars[which.identified,],pch="O",col=2,cex=2)

plot(rnorm(100),rexp(100,1),xlab=expression(hat(mu)[x]),
ylab=expression(eta^tau),
main=expression(paste("Scatterplot of ", eta^tau, "
versus ", hat(mu)[x])))
```

Identify “n” points by clicking

```
# PLOTS, CONT.
x=rnorm(100);y=x+rnorm(100)
plot(x,y,axes=F)
axis(1)
mtext("I can put text on this side",3)
mtext("I can put text on this side too",4,col=2)
text(0,1,"text1",col="blue")
text(.5,1.5,"text2",srt=90,col="green")
text(.5,-1.5,"text3",col="purple",cex=2)

?swiss
pairs(swiss)
cor(swiss)
swiss.pch=rep("X",dim(swiss)[1])
swiss.pch[swiss$Catholic>50]="O"
swiss.colors=rep("red",dim(swiss)[1])
swiss.colors[swiss$Education>median(swiss$Education)]="
blue"
pairs(swiss,pch=swiss.pch,col=swiss.colors)
```

Putting text on plot

Using “pairs” to explore two-way relationships

**# REGRESSION MODELS**

(http://people.musc.edu/~slateeh/Rintro/Rdemo.htm)

```
plot(faithful$eruptions, faithful$waiting, main = "Eruptions
of Old Faithful",
        xlab = "Eruption time (min)", ylab = "Waiting time
to next eruption (min)")
lines(lowess(faithful$eruptions, faithful$waiting, f = 2/3,
iter = 3),
           col = "red", lwd = 3)
fit = lm(waiting ~ eruptions, data = faithful)
abline(fit, col = "green", lwd = 3)
plot(fit, which = 1:3)
```

```
# SIMPLE SYMBOLIC DERIVATIVES AND NUMERIC INTEGRALS
D(expression( x^2), "x")
D(expression( x^2-exp(x)), "x")
D(expression( x^2-log(x)), "x")
D(D(expression(x^2),"x"),"x")
D(D(D(expression(x^2),"x"),"x"),"x")
D(expression( y*x^2), "x")

my.deriv=deriv(~x^2-log(x),"x")
x=1:5
eval(my.deriv)
my.deriv2=deriv(~x^2-log(x)+x*y^2,"x")
y=3:7
eval(my.deriv2)

integrate(dnorm, -1.96, 1.96)
integrate(dnorm, -Inf, Inf)
```

```
# CONDITIONS AND LOOPS ###

# R follows the form  "if (expr_1) expr_2 else expr_3"
# and "{" used to set apart grouped statements

x=1
if (x>1) print("BINGO") else print("BOINGO")
if (x>1) {print("BINGO")} else {print("BOINGO")}
if (x>1) {print("BINGO")} else {print("BOINGO");print("BOINGO
BOINGO")}

# "for" loops: "for (name in expr_1) expr_2"
for (i in (1:10)){
 print(i^2)
 }
# nested loops
for (i in (1:5)){
 for (j in (1:10)){
  print(c(i,j))
  }
 }
```

```
# nested loops
for (i in (1:5)){
 for (j in (1:10)){
  print(c(i,j))
  }
 }

for (i in (1:5)){
 for (j in (i:10)){
  print(c(i,j))
  }
 }
```

```
# Working with entire R objects is often much
faster than looping
start.time=Sys.time()
num.rows=1000
num.cols=1000
big.matrix=matrix(0,num.rows,num.cols)

start.time=Sys.time()
for (i in (1:num.rows)){
 for (j in (1:num.cols)){
  big.matrix[i,j]=i+j
  }
 }
end.time=Sys.time()
end.time-start.time

start.time=Sys.time()
big.matrix=outer(1:num.rows,1:num.cols,FUN="+"
)
end.time=Sys.time()
end.time-start.time
```

# Writing functions

```
find.quadratic=function(a,b,c){
 root1=(-b+sqrt(b^2-4*a*c))/(2*a)
 root2=(-b-sqrt(b^2-4*a*c))/(2*a)
 return(list(root1=root1,root2=root2))}

find.quadratic(a=2,b=10,c=3)
find.quadratic(2,10,3)

# Now let's do this completely numerically
error.quad=function(x,a,b,c){
 current=a*x^2+b*x+c
 error=(current-0)^2
 return(error)}

optimize(error.quad,a=2,b=10,c=3,interval=c(-
1,0))
optimize(error.quad,a=2,b=10,c=3,interval=c(-5,-
1))
help(optimize)
```

```
### THIS FUNCTION DOES LINEAR REGRESSION
regression=function(y,x){
 # y is a vector
 # x is the design matrix.
 n=length(y)
 p=dim(x)[2]
 xTx.inv=solve(t(x)%*%x)
 betahat=as.vector(xTx.inv%*%t(x)%*%y)
 resid=y-x%*%betahat
 varhat=(sum(resid^2))*diag(xTx.inv)/(n-p)
 se=sqrt(varhat)
 pval=2*pt(-abs(betahat/se),df=n-p)
 out=list(betahat,se,betahat/se,n-p,pval,resid)
 names(out)=c("betahat","se","t","df","pval","resid")
 return(out)
 }
```

# R Workspace

*# save your command history
savehistory(file="*myfile*") # default is ".Rhistory"

# recall your command history
loadhistory(file="*myfile*") # default is ".Rhistory“

# Graphics

- #Save plots

```
x<-rnorm(100)
y<-2*x+ rnorm(100)
plot(x,y)
pdf("test.pdf")
plot(x,y)
dev.off()
```

# Lecture 3: Simple Linear Regression

Reading

- Weisberg, Chapter 2: "Simple Linear Regression" (Required)

We will consider the case in which we observe a single response and one covariate. (For most of the course, we assume that the covariates are fixed and known.)

This lecture will serve as a gentle introduction to the matrix-based material utilized in the rest of this course, and provide some intuition and motivations for later concepts.

## Model Notation

- Roman letters $(a, b, c, \ldots, x, y, z)$ represent constants and random variables.
  - Letters at beginning of the alphabet, $(a, b, c, \ldots)$, often represent constants.
  - Letters in the middle of the alphabet, $(\ldots, i, j, k, \ldots)$, often represent indices.
  - Letters at the end of the alphabet, $(\ldots, x, y, z)$, often represent random variables.
- Greek letters $(\alpha, \beta, \gamma, \ldots, \chi, \psi, \omega)$ represent parameters to be estimated.
- Lower-case symbols in bold type $(\mathbf{x}, \mathbf{y}, \boldsymbol{\beta}, \ldots)$ represent vectors.
- Upper-case symbols in bold type $(\mathbf{X}, \mathbf{A}, \boldsymbol{\Sigma}, \ldots)$ represent matrices.

## Simple Linear Regression

Regression analysis is used to explore the nature of the relationship between a response variable and one or more covariates. In simple linear regression, we focus on the case when we consider only a single covariate.

*Hypothesis:* State-specific melanoma mortality rates are related to the latitude of the state.

Define:

$y_i$=annual melanoma mortality in state $i$, $i = 1, \ldots, 50$
$x_i$=latitude of the center of the state (in degrees)

Our observations are $(x_1, y_1), \ldots, (x_{50}, y_{50})$, where the $x_i$'s are known fixed values (*predictors*, *covariates*, or *independent variables*), and the $y_i$'s are random *response* variables (*dependent variables*).

| *State* | *Melanoma Mortality Rate* (Deaths/Million) | *Latitude* (in degrees) |
|---|---|---|
| Alabama | 219 | 33.00 |
| Alaska | 220 | 63.25 |
| Florida | 197 | 28.00 |
| Hawaii | 330 | 19.96 |
| Maine | 117 | 45.20 |
| Minnesota | 116 | 46.00 |
| Mississippi | 207 | 32.80 |
| North Dakota | 115 | 47.50 |
| North Carolina | 199 | 35.50 |
| Tennessee | 186 | 36.00 |
| Vermont | 153 | 44.00 |

How we do determine the relationship between melanoma mortality and latitude?

## Simple Linear Regression

Let us write the simple linear regression (SLR) model as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n \tag{1}$$

where

- $n$ is the number of observations or sampling units
- $y_1, \ldots, y_n$ are the random responses,
- $x_1, \ldots, x_n$ are the fixed and known covariates
- $\varepsilon_1, \ldots, \varepsilon_n$ is a vector of unobserved random errors, s.t. $\varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$
- $\beta_0, \beta_1$, and $\sigma^2$ are parameters (fixed and to be estimated from the data)

The index $i$ corresponds to subjects or sampling units.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, n$$

where

- $E[\varepsilon_i | x_i] = E[\varepsilon_i] = 0$
- On right side, only $\varepsilon_i$ is random
- $(y_i, x_i)$ forms the data for observation $i$

## Model Assumptions

- Homogeneity
- Independence
- Linearity
- Existence
- Gaussian Errors

Violations of these assumptions lead to problems with the application, estimation, and interpretation of linear models.

We also may write the simple linear regression (SLR) model by taking expectations of both sides

$$\begin{aligned} E[y_i|x_i] &= E[\beta_0 + \beta_1 x_i + \epsilon_i|x_i] \\ &= \beta_0 + \beta_1 x_i \\ Var[y_i|x_i] &= Var[\beta_0 + \beta_1 x_i + \epsilon_i|x_i] \\ &= \sigma^2 \end{aligned}$$

Based on the assumptions of the SLR, how did we arrive at this? What does this say about the distribution of $y_i$? We can plot this line on the next slide

How we interpret the parameters $\beta_0$, $\beta_1$, and $\sigma^2$ in this model?

- $\beta_0$: intercept, $E[y_i|x_i = 0]$
- $\beta_1$: slope, change in $E[y_i|x_i]$ when $x_i \uparrow$ by 1 unit
- $\sigma^2$: Variance of $y_i$ about $E[y_i|x_i]$

Lets say that we knew $\beta_0$, $\beta_1$, and $\sigma^2$, and had covariates $x_1, \ldots, x_n$. Then, we could generate new data $y_1, \ldots, y_n$ using eq. 1.

However, in reality we typically only have observed data points $(x_1, y_1), \ldots, (x_n, y_n)$ and do not know $\beta_0$, $\beta_1$, and $\sigma^2$, and do not observe $\varepsilon_1, \ldots, \varepsilon_n$. We estimate these equantities given the observed data. How to choose the best line?

Going back to the original example, we have the model:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, 50$,

where $(\beta_0, \beta_1)$ are *parameters* to be estimated. The $\varepsilon_i$'s represent random errors, which account for the fact that the response will vary even for states with the same latitudes (see, for example, MS and AL). We may wish to use this model to do the following.

- Estimate $\beta_0$, $\beta_1$, and $\sigma^2$.
- Test hypotheses about $\beta_1$ or obtain confidence limits for $\beta_1$, relating the statistical results back to the scientific question concerning the relationship between latitude and melanoma mortality.
- Predict a future $y$ at a given $x$.

Fitting the data, we obtain the estimates $\widehat{\beta_0}$ =389.189 and $\widehat{\beta_1}$ =-5.978.

We can use these estimates to draw the estimated regression line.

How do we interpret our parameter estimates?

## Ordinary Least Squares (OLS) Estimation

- Many possible values of $\beta_0$, $\beta_1$
- Many different possible lines
- OLS is an approach to obtain estimates of these parameters, $\hat{\beta}_0$, $\hat{\beta}_1$, in addition to $\hat{\sigma}^2$

With these estimated values we may obtain two values

$$\hat{y}_i = \widehat{E}[y_i|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$$

where $\hat{y}_i$ is the predicted value of $y_i$ given $\hat{\beta}_0$, $\hat{\beta}_1$, $x_i$, and

$$\hat{\varepsilon}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), i = 1, \ldots, n$$

where $\hat{\varepsilon}_i$ is called the residual for the $ith$ observation.

Intuition behind OLS - graphical illustration (Weisberg pg. 25). Solid line is the fitted line, dashed line is the true line (why are these different?) that the data was generated from. But how do we select the fitted line, and how do we know it is "best" given the observed data?

**Least Squares Criterion** The values for $(\hat{\beta}_0, \hat{\beta}_1)$ are obtained by minimizing the RSS with respect to $(\beta_0, \beta_1)$. That is, we minimize

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Define the Residual Sum of Squares (RSS), also known as the Sums of Squares for Error (SSE), as the following

$$\begin{aligned} RSS(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \\ &= S_{YY} - \hat{\beta}_1^2 S_{XX} \end{aligned}$$

where $S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ and $S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

But how do we minimize this $RSS(\beta_0, \beta_1)$? Discussed in Weisberg A.3, we will briefly go over approach here.

Intuitively, $\hat{\sigma}^2$ can be estimated by averaging $\hat{\varepsilon}_i^2$. An unbiased estimate for $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

This quantity is also referred to as the Residual Mean Square, or the Mean Square of the Errors (MSE). RSS may also be expressed as

$$RSS \quad = \quad \sum_{i=1}^{n} [\hat{\varepsilon}_i]^2$$

Why is this the case? Note that RSS is determined solely on $\hat{\beta}_0$ and $\hat{\beta}_1$.

Assuming that the errors are drawn from a normal distribution, then the Residual Mean Square divided by $\sigma^2$ will be distributed Chi-squared with $n-2$ degrees of freedom. Then, we have that

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2}\chi^2(n-2)$$

so that

$$\begin{aligned} E[\hat{\sigma}^2] &= \frac{\sigma^2}{n-2}E[\chi^2(n-2)] \\ &= \frac{\sigma^2}{n-2}(n-2) \\ &= \sigma^2 \end{aligned}$$

$\rightarrow \hat{\sigma}^2$ is unbiased. This is in contrast to the MLE of $\sigma^2$.

**Properties of Least Squares Estimates** We can see that $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the random $\varepsilon_i$'s (why?). If the errors are mean 0, then

$$\begin{aligned} E[\hat{\beta}_0] &= \beta_0 \\ E[\hat{\beta}_1] &= \beta_1 \end{aligned}$$

In other words, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased. Notice that we do not need the assumption of normality for this result to hold. The normal assumption of the errors impacts hypothesis testing.

$$
\begin{aligned}
E[\hat{\beta}_0] &= \beta_0 \\
E[\hat{\beta}_1] &= \beta_1
\end{aligned}
$$

Can you show how these estimators are unbiased?

Now lets look at the variances of teh estimates

$$
\begin{aligned}
Var[\hat{\beta}_0] &= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \\
Var[\hat{\beta}_1] &= \frac{\sigma^2}{S_{XX}}
\end{aligned}
$$

How do we obtain this result?

We can also derive the forms of the covariances and correlation of the estimates

$$
\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= -\sigma^2 \frac{\bar{x}}{S_{XX}} \\
\rho(\hat{\beta}_0, \hat{\beta}_1) &= \frac{-\bar{x}}{\sqrt{\frac{S_{XX}}{n} + \bar{x}^2}}
\end{aligned}
$$

What do these results imply? How do we obtain this result?

The Gauss-Markov theorem provides an optimality result for ols estimates. Among all estimates that are linear combinations of the $y_1, \ldots, y_n$ and unbiased, the ols estimates have the smallest variance. These estimates are called the best linear unbiased estimates , or blue. If one believes the model assumptions and is interested in using linear unbiased estimates, the ols estimates are the ones to use.

The means and variances, and covariances of the estimated regression coefficients do not require a distributional assumption concerning the errors. Since the estimates are linear combinations of the $y_1, \ldots, y_n$ , and hence linear combinations of the errors, the central limit theorem shows that the coefficient estimates will be approximately normally distributed if the sample size is large enough.

For smaller samples, if the errors are i.i.d, then the regression estimates $(\hat{\beta}_0, \hat{\beta}_1)$ will have a joint normal distribution with means, variances, and covariances as given before. When the errors are normally distributed, the OLS estimates can be justified using a completely different argument, since they are then also maximum likelihood estimates.

## Estimated Variances

$$\begin{aligned}
\widehat{Var}[\hat{\beta}_0] &= \hat{\sigma}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{XX}} \right) \\
\widehat{Var}[\hat{\beta}_1] &= \frac{\hat{\sigma}^2}{S_{XX}} \\
se\left(\hat{\beta}_j\right) &= \sqrt{\widehat{Var}[\hat{\beta}_j]}, j = 1, 2
\end{aligned}$$

Standard deviation - variability of an observable random variable like the response $y_i$ or an unobservable random variance like the errors $\varepsilon_i$ .
Standard error - square root of the estimated variance of a statistic like $\hat{\beta}_1$ .

**Confidence Intervals and t-tests** Confidence intervals result in interval estimates, while tests provide methodology for making decisions concerning the value of a parameter or fitted value.

When the errors are $\overset{i.i.d}{\sim} N(0, \sigma^2)$, parameter estimates, fitted values, and predictions will be normally distributed. Why is this the case?

Confidence intervals and tests can be based on a t-distribution, which is the appropriate distribution with normal estimates but using $\hat{\sigma}^2$ to estimate the unknown variance $\sigma^2$. Suppose we let $t(\alpha/2, d)$ be the value that cuts off $\alpha/2 \times 100\%$ in the upper tail of the t-distribution with $d$ df.

For example, the standard error of the intercept is $se(\hat{\beta}_0) = \hat{\sigma}(1/n + \bar{x}^2/S_{XX})$. Hence, a $(1-\alpha) \times 100\%$ CI is the set of points $\beta_0$in the interval

$$\hat{\beta}_0 - t(\alpha/2, n-2)se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 - t(\alpha/2, n-2)se(\hat{\beta}_0).$$

$1-\alpha$ percent of such intervals will include the true value.

A hypothesis test of

$$H_0 : \beta_0 = \beta_0^*, \beta_1 \texttt{arbitrary},$$
$$H_A : \beta_0 \neq \beta_0^*, \beta_1 \texttt{arbitrary}$$

is obtained by computing the t-statistic $t = \frac{\hat{\beta}_0 - \beta_0^*}{se(\hat{\beta}_0)}$ and referring this ratio to the t-distribution with df $= n - 2$, the number of df in the estimate of $\sigma^2$.

From our Melanoma example, testing

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

would be addressing what question in particular?

**Prediction** If we found a new state and latitude value, $x_{51}$, can we predict what the mortality would be, given the fitted model? That is, we would like to know what $y_{51}$ would be given $x_{51}$, assuming the existing data is relevant to the new data. A point prediction of $y_{51}$, say $\tilde{y}_{51}$, is simply

$$\tilde{y}_{51} = \hat{\beta}_0 + \hat{\beta}_1 x_{51}$$

where $\tilde{y}_{51}$ predicts the unobserved $y_{51}$.

Assuming the model is correct, then the true value of $y_{51}$ is

$$\tilde{y}_{51} = \beta_0 + \beta_1 x_{51} + \varepsilon_{51}$$

where $\varepsilon_{51}$ is the random error for attached to the future value. Since $\tilde{y}_{51}$ is based on estimated values of $\beta_0$ and $\beta_1$, the prediction error variability will have a second component tha arises from the uncertainty in the estimtates of the coefficients. We can show that

$$Var(\tilde{y}_{51}|x_{51}) = \sigma^2 + \sigma^2 \left( \frac{1}{n} + \frac{(x_{51} - \bar{x})^2}{S_{XX}} \right).$$

What does this imply?

## Coefficient of Determination $R^2$

Ignoring all covariates, the best prediction of the response would be the total sum of squares $SSY = \sum i = 1^n (y_i - \bar{y})^2$, the observed total variation of the response, ignoring any and all predictors.

The unexplained variation is given by $RSS$, the sum of squared deviations from the fitted line. We define $SSreg = SSY - RSS$. If both sides are divided by $S_{YY}$, then we have $\frac{SSreg}{S_{YY}} = 1 - \frac{RSS}{S_{YY}} = R^2$. $R^2$, the coefficient of determination, may be thought of the proportion of the total variability explained by the model, or 1 - the unexplained variability. This is a scale-free one-number summary of the strength of the relationship between $x_i$ and $y_i$. In later lectures we will go more into this summary.

## Residuals

Plots of residuals versus other quantities are used to find failures of assumptions. The most common plot, especially useful in simple regression, is the plot of residuals versus the fitted values. A null plot would indicate no failure of assumptions. Curvature might indicate that the fitted mean function is inappropriate. Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance. A few relatively large residuals may be indicative of outliers, cases for which the model is somehow inappropriate.

Residuals
5
0
−5
60
62
64
66
68
Fitted values

In a later lecture, we will learn how to use residuals to check model assumptions.

## SLR in R and SAS

Lets take a peek at the data for the contiguous US

| | mortality | latitude | longitude | ocean |
|---|---|---|---|---|
| Alabama | 219 | 33.0 | 87.0 | yes |
| Arizona | 160 | 34.5 | 112.0 | no |
| Arkansas | 170 | 35.0 | 92.5 | no |
| California | 182 | 37.5 | 119.5 | yes |
| Colorado | 149 | 39.0 | 105.5 | no |
| Connecticut | 159 | 41.8 | 72.8 | yes |

Lets plot the data

```
> plot(USmelanoma[,2], USmelanoma[,1],
+      ylab = "Mortality (Deaths per Million)",
+      xlab = "Latitude"
+ )
```

Mortality (Deaths per Million)
100
140
180
220
30
35
40
45
Latitude

We can fit the linear regression model with the `lm` function

```
> out = lm(mortality ~ latitude, data = USmelanoma)
> out = lm(USmelanoma[,1] ~ USmelanoma[,2])
```

The summary function prints basic information about the fit

```
> summary(out)

Call:
lm(formula = USmelanoma[, 1] ~ USmelanoma[, 2])

Residuals:
    Min      1Q  Median      3Q     Max
-38.485 -12.823   1.272  12.192  44.381

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      388.312     23.767  16.338  < 2e-16 ***
USmelanoma[, 2]   -5.966      0.597  -9.994 4.15e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 19.07 on 46 degrees of freedom
Multiple R-squared:  0.6847,        Adjusted R-squared:  0.6778
F-statistic: 99.89 on 1 and 46 DF,  p-value: 4.145e-13
```

We can also extract values based on the model fit

```
> residuals = out$residuals
> fitted = out$fitted.values
> coefficients = out$coefficients
```

```
> plot(fitted, residuals)
```

Now, lets use what we learned so far to check the results

```
> y = USmelanoma$mortality
> x = USmelanoma$latitude
> n = length(y)
> ybar = mean(y)
> xbar = mean(x)
> S_{YY} = sum( (y - ybar)^2 )
> S_{XX} = sum( (x - xbar)^2 )
> S_{XY} = sum( (y - ybar)*(x - xbar) )
```

Now, lets use what we learned so far to check the results

```
> coefficients

    (Intercept) USmelanoma[, 2]
     388.311830       -5.966476

> S_{XY}/S_{XX} # beta_1 hat

[1] -5.966476

> ybar - (S_{XY}/S_{XX})*xbar  # beta_0 hat

[1] 388.3118
```

```
> RSS = sum( (y - fitted)^2 )
> s2_hat = RSS/(n - 2) # sigma_squared hat
> s2_hat

[1] 363.5953

> sqrt(s2_hat*(1/n + xbar^2/S_{XX})) # se of beta_0 hat

[1] 23.76711

> sqrt(s2_hat/S_{XX}) # se of beta_1 hat

[1] 0.5969898

> round(vcov(out),3) # covariance matrix of beta_0 hat, beta_1 hat

                (Intercept) USmelanoma[, 2]
(Intercept)         564.875         -14.093
USmelanoma[, 2]     -14.093           0.356
```

From our Melanoma example, testing

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

$t = \frac{-5.966-0}{0.597} = -9.994$

```
> t = -9.994
> p.value  = 2*(1 - pt(abs(t), n-2))
> p.value
[1] 4.147793e-13
```

## SAS Proc Reg

```
proc reg data = USmelanoma;
  model mortality = latitude;
run;
```

## The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: mortality

| | |
|---|---|
| Number of Observations Read | 49 |
| Number of Observations Used | 49 |

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 36464 | 36464 | 99.80 | <.0001 |
| Error | 47 | 17173 | 365.38436 | | |
| Corrected Total | 48 | 53637 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 19.11503 | R-Square | 0.6798 |
| Dependent Mean | 152.87755 | Adj R-Sq | 0.6730 |
| Coeff Var | 12.50349 | | |

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | 1 | 389.18935 | 23.81232 | 16.34 | <.0001 |
| latitude | 1 | -5.97764 | 0.59837 | -9.99 | <.0001 |

## Mini Problems to do from Weisberg

- 2.1.3
- 2.11

# Lecture 5: General Linear Model: Estimation and Testing

*Reading Assignment:*

- Muller and Fetterman Chapter 2
- Weisberg Chapter 3

We will now consider the general case in which we observe a single response and one or more covariates.

We write the general form of the linear model as

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\, \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

where

- $\mathbf{y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{bmatrix}$ contains the observed responses,

- $\mathbf{X}_{n\times p} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{p-1}) = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,p-1} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & & & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix}$ is a matrix of fixed and known covariates,

- $\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})' = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ is a vector of parameters to be estimated, and

- $\boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is a vector of unobserved random errors.

- The rows of $\mathbf{y}, \mathbf{X},$ and $\boldsymbol{\varepsilon}$ correspond to subjects or sampling units.
- Columns of $\mathbf{X}$ and the corresponding rows of $\boldsymbol{\beta}$ correspond to predictors. Often, the first column of $\mathbf{X}$, denoted $\mathbf{x}_0$, corresponds to an intercept variable and takes the value 1 for all subjects so that we have $\mathbf{x}_0 = \mathbf{J} = \mathbf{1}$. (We will assume this to be the case unless we state otherwise.)

We refer to the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as the matrix representation of the linear model. All linear models may be written in this form.

Alternatively, the model is written in scalar notation as

$$y_i = \sum_{j=0}^{p-1} x_{ij}\beta_j + \varepsilon_i, \ i = 1, \ldots, n.$$

The row of $\mathbf{X}$ corresponding to subject $i$ is denoted $\text{row}_i(\mathbf{X})$ or $\mathbf{x}_i'$ (which may lead to some confusion with the columns of $\mathbf{X}$ so that the precise meaning may depend on the context).

## Example: Social Setting, Family Planning, and Birth Rate

Rodriguez (2002) considers factors related to the decline in the crude birth rate (CBR, the number of births per thousand population) between 1965 and 1975 for 20 Latin American and Caribbean countries. Covariates of interest include a social setting index (SOCIAL) and a family planning index (FAMPLAN). The social setting index is a function of literacy, school enrollment, life expectancy, infant mortality, percent of males aged 15-64 in the non-agricultural labor force, gross national product per capita, and percent of population living in urban areas. Higher social setting scores represent higher socio-economic levels. The family planning index is a function of availability of contraceptive methods, official government family planning policies, and structure of family planning programs in the country. Values of 20 or more indicate strong efforts in family planning, and values of 10-19 represent moderate efforts.

A few representative observations in the birth rate dataset are presented below.

| *Country* | *SOCIAL* | *FAMPLAN* | *CBR* |
|---|---|---|---|
| Brazil | 74 | 0 | 10 |
| Costa Rica | 84 | 21 | 29 |
| Haiti | 35 | 3 | 0 |
| Mexico | 83 | 4 | 9 |
| Trinidad-Tobago | 84 | 15 | 29 |

This model is written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as follows:

$$\mathbf{y}_{20\times 1} = \mathbf{X}_{20\times 3}\,\boldsymbol{\beta}_{3\times 1} + \boldsymbol{\varepsilon}_{20\times 1}$$

$$\begin{bmatrix} 10 \\ 29 \\ 0 \\ 9 \\ 29 \\ \vdots \end{bmatrix}_{20\times 1} = \begin{bmatrix} 1 & 74 & 0 \\ 1 & 84 & 21 \\ 1 & 35 & 3 \\ 1 & 83 & 4 \\ 1 & 84 & 15 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{20\times 3} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{20\times 1}$$

- $\beta_0$ is called the intercept, which is the expected decline in birth rate when the social setting and family planning indices take the value zero.
  - Some investigators prefer to center all predictors so that the mean of each predictor is zero. Why?
- $\beta_1$ is the slope for social setting. It is interpreted as the expected increase in CBR decline for a one unit increase in the social setting index.
- $\beta_2$ is the slope for family planning. It is interpreted as the expected increase in CBR percent decline for a one unit increase in the family planning index.
- Each element of $\varepsilon$ represents the distance between a country's observed percent CBR decline and the population regression line.

## Least Squares Estimation

The **least squares** estimate of $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}$, satisfies

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) &= \min_{\beta}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \min_{\beta} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 \\ &= \min_{\beta} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2,\end{aligned}$$

where $\mathbf{x}_i'$ indicates the $i$-th row of matrix $\mathbf{X}$. The least squares estimate of $\boldsymbol{\beta}$ minimizes the squared Euclidean distance between $\mathbf{y}$ and its mean $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

It can be shown that when we make the additional assumption that $y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$, the least squares estimates $\widehat{\boldsymbol{\beta}}$ are also *maximum likelihood* estimates. Maximum likelihood estimates have several desirable properties that you will learn more about in BIOS 661.

Below is a plot of the observed percent CBR decline versus family planning effort along with the least squares regression line. Consider two steps in fitting this line. First, we calculate the mean percent CBR decline, $\overline{y}$. Then, we pivot a line around $(\overline{x}, \overline{y}) = (9.55, 14.3)$, where $\overline{x}$ is the mean family planning effort, until we minimize the sum of squared deviations around the line.

CBR Decline by Family Planning Effort
Percent CBR Decline
0
10
20
30
40
0
5
10
15
20
Family Planning Effort

The least squares estimator, $\widehat{\boldsymbol{\beta}}$, has several good properties.

- First, if the linear model assumption holds, then the least squares estimator is *unbiased*; that is, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

- Next, we will show later that if the observations are uncorrelated and have constant variance $\sigma^2$, then the variance-covariance matrix of the least squares estimator $\widehat{\boldsymbol{\beta}}$ is $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. This estimator is *efficient* in the sense that it has the smallest variance in the class of all unbiased estimators that are linear functions of the data. We call this type of estimator a BLUE (Best Linear Unbiased Estimator).

- If we add the assumption that the error vector is normally distributed, then the least squares estimator is the "best" estimator among *all* unbiased estimators. (We define a "best" estimate to be an unbiased estimate with minimum variance. Other criteria for "best" estimates do exist but will not be addressed further here.)

- Later, we will also show that the sampling distribution of the least squares estimator $\widehat{\boldsymbol{\beta}}$ in *large* samples is approximately multivariate normal with the previously given covariance; that is, $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. In the intercept-only model, this means that $\widehat{\boldsymbol{\beta}} = \bar{\mathrm{y}} \sim N(\mu, \frac{\sigma^2}{n})$ in large samples. (This is true even if your errors are not exactly normal but are uncorrelated with constant variance. If they are exactly normal, then the sampling distribution of $\widehat{\boldsymbol{\beta}}$ is exactly normal, and you don't need to worry about attaining a certain sample size in order for asymptotic theory to hold.)

Least squares estimation requires a variety of assumptions:

**Existence Assumption**

Assume $\varepsilon_i$ has finite first and second moments. That is, we observe values of random variables with finite variance. In practice, considering a finite number of subjects ensures that this assumption holds.

**Linearity Assumption**

We assume the expected values (means) of the response are linear functions of the parameters. That is, we assume

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta},$$

or equivalently, that

$$E(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta} = \mathbf{0} \Leftrightarrow \mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

The linearity assumption refers to the relationship between the response and parameters. Consider several examples.

1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is linear in parameters, predictors, and error
2. $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is linear in parameters and error
3. $y_i = \beta_0 + x_i^{\beta_1} + \varepsilon_i$ is linear in error
4. $y_i = \beta_0 \exp(-\beta_1 x_i) + \varepsilon_i$ is nonlinear in parameters
5. $y_i = \beta_0 \left[\exp(-\beta_1 x_i)\right] \varepsilon_i \Leftrightarrow \text{log}(y_i) = \text{log}(\beta_0) - \beta_1 x_i + \text{ln}(\varepsilon_i)$
   $\Leftrightarrow$ (5b)${y_i}^* = \gamma_0 + \gamma_1 x_i + \delta_i$, which is linear in parameters, predictor, and error
6. $y_i = (\beta_0 + \beta_1 x_i)\varepsilon_i$ is nonlinear in parameters

Of these examples, only 1, 2, and 5b meet the linear model assumption of linearity. If linearity does not hold, then we should not attempt to fit linear models, and estimates from linear models will not have the nice properties discussed previously. Nonlinear models, such as the exponential growth or decay model given by $\mathbf{y} = \beta_1 e^{\beta_2 \mathbf{x}} + \varepsilon$, will not be covered in this class.

## Independence Assumption

Each element of $\varepsilon$ is statistically independent of every other. Equivalently, each element of $\mathbf{y}$ is statistically independent of every other, conditional on $\mathbf{X}$. This generally is fairly obvious to check.

For example, if a set of twins or a parent and child are included in data, the independence assumption will be violated. If positively correlated observations tend to have positively correlated predictor values, standard linear regression will yield anti-conservative tests of associations (i.e., your p-value will be too small). This is the type of error we expect with correlation in family data or longitudinal data.

We will discuss appropriate models for correlated data later in BIOS 663.

## Homogeneity Assumption

We assume each element of $\boldsymbol{\varepsilon}$ has the same variance $\sigma^2$. Equivalently, $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$. It is possible to check this assumption, and we will spend a good deal of time later in the course discussing ways to check homogeneity of variances (which is sometimes called **homoscedasticity**). *Discussion of Homogeneity:*

- $\sigma_i^2$ is the variance of the error for subject $i$.
- $\sigma_i^2 = \sigma_i^2(y_i \mid x_{i1}, x_{i2}, \ldots, x_{ip})$ is the variance of the response, conditional on the value of the covariates $\mathbf{x}_i$ for subject $i$.
- Homogeneity of variances means that $\sigma_i^2 = \sigma^2$ for all subjects $i$.
- Homogeneity also means that the variance about the regression function $E(\mathbf{y} \mid \mathbf{X})$ is constant.

*Error Covariance Matrix*

Recall that we assume errors have mean zero. So $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. Thus

$$\begin{aligned}
\text{Cov}(\boldsymbol{\varepsilon}) &= E\left[(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))'\right] \\
&= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') - \boldsymbol{\varepsilon}E(\boldsymbol{\varepsilon})' - E(\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}' + E(E(\boldsymbol{\varepsilon})E(\boldsymbol{\varepsilon})') \\
&= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') - \mathbf{0} - \mathbf{0} + \mathbf{0} = \mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')_{\mathbf{n}\times\mathbf{n}} \\
&= \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2)\ldots & E(\varepsilon_1\varepsilon_n) & \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & & \vdots \\ \vdots & & & \vdots \\ E(\varepsilon_n\varepsilon_1) & \ldots & \ldots & E(\varepsilon_n^2) \end{bmatrix}.
\end{aligned}$$

If the covariance $E(\varepsilon_i\varepsilon_j) = 0$, then the correlation

$$\frac{E(\varepsilon_i\varepsilon_j)}{\sqrt{E(\varepsilon_i^2)}\sqrt{E(\varepsilon_j^2)}} = 0,$$

and vice versa. Although independence of two random observations

always implies zero correlation (and zero covariance), the converse is not true. That is, unless we have Gaussian random variables, zero covariance does *not* imply independence.

The independence assumption means that the error covariance matrix is a diagonal matrix. In addition, because the covariance of a variable with itself equals its variance, we have

$$\begin{aligned}
\mathsf{Cov}(\boldsymbol{\varepsilon}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\
&= \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & \sigma_n^2 \end{bmatrix}.
\end{aligned}$$

The homogeneity assumption implies that $\sigma_i^2 = \sigma^2$ so that

$$\begin{aligned}
\text{Cov}(\boldsymbol{\varepsilon}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\
&= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n.
\end{aligned}$$

For example, if the CBR decline is more variable for countries with low family planning effort than for countries with high family planning effort, conditional on other covariates of interest, the homogeneity of variances assumption would not hold.

## Gaussian Errors Assumption

The Gaussian errors assumption is that $\varepsilon_i \sim N(0, \sigma_i^2)$. Assuming homogeneity of variances as well, we have that $\varepsilon_i \sim N(0, \sigma^2)$. This implies that $y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$ so that

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)^2\right).$$

The normality assumption is not needed for validity of least squares estimation. However, adding the assumption of Gaussian errors means that the least squares estimate of $\boldsymbol{\beta}$ is also a maximum likelihood estimate and a minimum variance unbiased estimate. The assumption of Gaussian errors also allows simple construction of exact small-sample hypothesis tests.

We combine all five assumptions as follows:

$$y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2),$$

with $y_i \perp y_j$ for $i \neq j$. Muller and Fetterman use the mnemonic *HILE Gauss* to describe the five assumptions.

Under the five assumptions (HILE Gauss), the $\{\varepsilon_i\}$ are i.i.d.(independent and identically distributed). However, the $\{y_i\}$ in general are not i.i.d.as $E(y_i) \neq E(y_j)$ because $\mathbf{x}_i' \neq \mathbf{x}_j'$ for all $i \neq j$. The $\{y_i\}$ are independent Gaussian random variables with equal variances, but they do not necessarily have equal expected values.

## Defining the Normal Equations

For $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, there are two distinct sets of parameters:

- $\boldsymbol{\beta}$ with $p$ elements, and
- $\sigma^2$ with 1 element.

We almost always assume $n \gg p$ so that the sample size $\gg$ the number of parameters.

Once the parameters $\boldsymbol{\beta}$ have been estimated, one may compute $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, the *predicted values* of the outcome. The *residuals* measure the accuracy of the prediction and are given by

$$\mathbf{y} - \widehat{\mathbf{y}} = \widehat{\boldsymbol{\varepsilon}} \neq \boldsymbol{\varepsilon}.$$

The error, $\varepsilon_i$, is unobserved, while the residual, $\widehat{\varepsilon}_i$, is observed (and an estimate of $\varepsilon_i$).

We seek estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ that are optimal in some sense. One criterion is least squares, which computes $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ that minimize the average squared distance from observed outcomes to predicted outcomes.

Total squared error of prediction, called the residual $SS$ or the sum of squares for error ($SSE$), is

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

To find $\widehat{\boldsymbol{\beta}}$ that minimizes $SSE$ take derivatives:

$$\frac{\partial SSE}{\partial \boldsymbol{\beta}} = \mathbf{0} - 2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

Set $\partial SSE/\partial \boldsymbol{\beta} = \mathbf{0_{p \times 1}}$ to create a system of $p$ simultaneous equations:

$$\mathbf{0} \ = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Rearranging terms yields the *normal equations*:

$$(\mathbf{X}'\mathbf{X})_{p \times p} \ \widehat{\boldsymbol{\beta}}_{p \times 1} = \mathbf{X}'_{p \times n} \ \mathbf{y}_{n \times 1}.$$

A $\widehat{\boldsymbol{\beta}}$ that solves the normal equations yields predicted values $\widehat{\mathbf{y}}$ and residuals $\widehat{\boldsymbol{\varepsilon}}$ such that $\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{\varepsilon}} = 0$; that is, the predicted values and residuals are orthogonal (normal to each other).

## Solving Normal Equations if $\mathbf{X}$ is Full Rank

If $\boldsymbol{X}'\boldsymbol{X}$ is full rank (i.e.,rank$(\mathbf{X}) = p$ with $n > p$), then $\mathbf{X}'\mathbf{X}$ has a unique inverse, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then we solve the normal equations:

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.
\end{aligned}$$

*Facts about the Least Squares Estimate:*

- $\widehat{\boldsymbol{\beta}}$ is the unique least squares estimate.
- $\widehat{\boldsymbol{\beta}}$ is best linear unbiased estimate (BLUE).
- With Gaussian errors, $\widehat{\boldsymbol{\beta}}$ is the MLE and minimum variance unbiased estimator.

## Solving Normal Equations for $\boldsymbol{X}$ Less Than Full Rank

If $\boldsymbol{X}$ is less than full rank, say rank$(\boldsymbol{X}) = r < p$, then $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist.

Solutions:

1. Drop some covariates.
2. Generalized Inverse ($\hat{\boldsymbol{\beta}}$ not unique)
3. Penalized regression methods, such as Ridge regression, which minimizes

$$\min_{\beta} \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

   where $\lambda$ is a tuning parameter that can be decided by cross-validation. In other words, we penalize the size of the regression coefficients The solution is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda I_{p\times p})^{-1}\mathbf{X}'\mathbf{y}$

**Hat Matrix** From now on, we assume $\mathbf{X}$ is full rank, unless otherwise specified.

The *hat matrix* is

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

The hat matrix has the special properties of idempotency ($\mathbf{HH} = \mathbf{H}$) and symmetry. In addition, rank$(\mathbf{H}) = r =$ rank$(\mathbf{X})$, and rank$(\mathbf{I} - \mathbf{H}) = n - r$. The prediction value can be written as

$$\widehat{\mathbf{y}} \quad = \quad \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y},$$

$\mathbf{H}$ is called the hat matrix because $\mathbf{H}\mathbf{y} = \widehat{\mathbf{y}}$. The residuals

$$\widehat{\boldsymbol{\varepsilon}} \quad = \quad \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

## Estimating $\sigma^2$

The MLE of $\sigma^2$ is $\frac{SSE}{n} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n}$.

$E\left[\frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n}\right] = \sigma^2\left[\frac{n-r}{n}\right]$, so the MLE of $\sigma^2$ is biased.

*Recall: Calculating Sample Variance*

When we calculate a variance, first we calculate a mean, say $\overline{y}$, and then we find the distance of each point from the mean, $y_i - \overline{y}$, $i = 1, \ldots, n$. (Note: by definition of the mean, $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$; for this reason, raw deviations are not a useful variance measure.) The *sum of squares* $\sum_{i=1}^{n}(y_i - \overline{y})^2$ is a useful measure of variability because it increases as the data are more dispersed about the mean. However, this measure also depends on $n$, and therefore it is not as useful in comparing groups. For comparison purposes, we convert the *sum of squares* to a *variance* by dividing by $n - 1$, where $n$ is the number of subjects in a group.

*Why not divide the sum of squares by $n$?*

The reason we do not divide by $n$ is that we do not have $n$ *independent* pieces of information about the variance. First, we calculated a mean, and then we calculated deviations from the mean. If we calculate the first $n-1$ deviations, then we know the last $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$. The independent pieces of information contributing to a statistic are called the *degrees of freedom*.

By similar logic, $\widehat{\boldsymbol{\sigma}}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-r} = \frac{SSE}{n-r}$ is an unbiased estimate of $\sigma^2$.

$\widehat{\sigma}^2$ is a *quadratic form* in $\boldsymbol{y}$:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-r} = \frac{(\mathbf{y}-\widehat{\mathbf{y}})'(\mathbf{y}-\widehat{\mathbf{y}})}{n-r} = \frac{\mathbf{y}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}}{n-r} \\
&= \frac{\mathbf{y}'\left[\mathbf{I}-\mathbf{H}\right]\mathbf{y}}{n-r}
\end{aligned}$$

## Example: Ozone Exposure Assessment

One common problem in environmental epidemiology is determining personal exposures to environmental toxicants, such as ozone in the air. Adverse health effects associated with ozone exposure include increased incidence of cough, chest pain, and other respiratory symptoms. Although outdoor ozone concentrations are monitored by the Environmental Protection Agency (EPA), it is more difficult to determine indoor concentrations. Personal exposures, which vary based on the proportion of time spent outdoors, at home, in the workplace, and in other areas, are even more difficult to measure. Using outdoor ozone concentrations as a crude approximation of personal exposure can lead to substantial measurement error, which can in turn lead to biased parameter estimates.

We consider data from a study conducted in State College, Pennsylvania, in which children wore small (2 cm $\times$ 3 cm) personal

ozone samplers. Investigators wish to model personal ozone exposures ($O_{PERSONAL}$) measured by the samplers as a function of outdoor ($O_{OUTDOOR}$) ozone concentrations (measured at a central State College site), home indoor ozone concentrations ($O_{HOME}$) for each child, and the proportion of time each child spent outdoors ($TIME_{OUTDOORS}$).

The data we consider include 64 measurements of personal ozone exposure (in parts per billion or ppb) along with the corresponding measurements of outdoor ozone concentrations, home indoor ozone concentrations, and the proportion of time spent outdoors.

This model is written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as follows:

$$\mathbf{y}_{64\times 1} = \mathbf{X}_{64\times 4}\,\boldsymbol{\beta}_{4\times 1} + \boldsymbol{\varepsilon}_{64\times 1}$$

$$\begin{bmatrix} 26.29 \\ 3.30 \\ 29.28 \\ 28.55 \\ 38.28 \\ \vdots \end{bmatrix}_{64\times 1} = \begin{bmatrix} 1 & 35.88 & 22.29 & 0.57 \\ 1 & 34.37 & 22.27 & 0.17 \\ 1 & 45.96 & 23.40 & 0.00 \\ 1 & 92.56 & 7.14 & 0.26 \\ 1 & 30.44 & 35.38 & 0.69 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{64\times 4} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{64\times 1}$$

- $\beta_0$ is called the intercept, which is the expected value of $O_{PERSONAL}$ when all other predictors

$$(O_{OUTDOOR}, O_{HOME}, TIME_{OUTDOORS})$$

  take the value zero.

- $\beta_1$ is the slope for outdoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in outdoor ozone concentration.

- $\beta_2$ is the slope for home indoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in home indoor ozone concentration.

- $\beta_3$ is the slope for the proportion of time spent outdoors. It is interpreted as the expected ppb increase in personal exposure for an additional one percent of time spent outdoors.

The following R code may be used to obtain parameter estimates for the ozone data.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

  (Intercept)  outdoor  home time_out
1           1 35.87771 22.29     0.57
2           1 43.79189 13.97     0.90
3           1 49.81255 18.96     0.55
4           1 34.37366 22.27     0.17
5           1 45.95496 23.40     0.00
6           1 64.76558 39.62     0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
```

The output from the R is given below.

```
> print(bhat)

                   [,1]
(Intercept)  3.78348593
outdoor      0.09142005
home         0.59543659
time_out    13.64453832

> print(mse)

         [,1]
[1,] 169.1365
```

The same estimates may be obtained from SAS PROC REG.

```
proc reg data=ozone;
model personal=outdoor home time_out;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: personal Personal Ozone Exposure (ppb)

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 5034.90667 | 1678.30222 | 9.92 | <.0001 |
| Error | 60 | 10148 | 169.13652 | | |
| Corrected Total | 63 | 15183 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 13.00525 | R-Square | 0.3316 |
| Dependent Mean | 23.54578 | Adj R-Sq | 0.2982 |
| Coeff Var | 55.23389 | | |

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value |
|---|---|---|---|---|---|
| Intercept | Intercept | 1 | 3.78349 | 4.34206 | 0.87 |
| outdoor | Outdoor Ozone Concentration (ppb) | 1 | 0.09142 | 0.09042 | 1.01 |
| home | Home Indoor Ozone Concentration (ppb) | 1 | 0.59544 | 0.16478 | 3.61 |
| time_out | Proportion of Time Spent Outdoors | 1 | 13.64454 | 7.70973 | 1.77 |

Parameter Estimates

| Variable | Label | DF | Pr > \|t\| |
|---|---|---|---|
| Intercept | Intercept | 1 | 0.3870 |
| outdoor | Outdoor Ozone Concentration (ppb) | 1 | 0.3160 |
| home | Home Indoor Ozone Concentration (ppb) | 1 | 0.0006 |
| time_out | Proportion of Time Spent Outdoors | 1 | 0.0818 |

## Hypothesis Testing

### The General Linear (Univariate) Hypothesis, GLH

For testing, we assume i.i.d.Gaussian errors. $\boldsymbol{\beta}$ is the matrix of primary parameters, and $\boldsymbol{\theta}_{a\times 1} = \mathbf{C}_{a\times p}\ \boldsymbol{\beta}_{p\times 1}$ is a matrix of secondary parameters, defined by $\mathbf{C}$, the *contrast matrix*. Each row of $\mathbf{C}$ defines a new scalar parameter in terms of the $\boldsymbol{\beta}$'s, e.g.,$\beta_1 - \beta_2$.

Let $\boldsymbol{\theta}_0$ be a matrix of known constants (the hypothesized values). Most often $\boldsymbol{\theta}_0$ is taken to be the zero matrix. The (univariate) general linear hypothesis is

$$\begin{aligned} H_0 : \boldsymbol{\theta}_{a\times 1} &= \boldsymbol{\theta}_0 \\ H_A : \boldsymbol{\theta}_{a\times 1} &\neq \boldsymbol{\theta}_0. \end{aligned}$$

## Example: Choosing Contrast Matrix and Secondary Parameter Matrix

For the ozone data, consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $O_{PERSONAL}$ as the response and $O_{OUTDOOR}$, $O_{HOME}$, and $TIME_{OUTDOORS}$ as predictors. Suppose we hypothesize that the outdoor and home exposures have the same effect on personal exposure levels. Thus we have $H_0 : \beta_1 - \beta_2 = 0$. This implies the secondary parameter $\theta = (\beta_1 - \beta_2)$ with corresponding contrast matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$ and $\theta_0 = 0$.

Suppose our hypothesis is that all the slopes are the same. What are the values of $\boldsymbol{\theta}$, $\mathbf{C}$, and $\boldsymbol{\theta}_0$?

## Estimability of a Parameter

Choosing appropriate $\boldsymbol{C}$ and $\boldsymbol{\theta}_0$, coupled with fitting appropriate models, allows testing hypotheses about all *estimable* parameters.

(Searle, 1971, p. 180): "A (linear) function of the parameters is defined to be estimable if it is identically equal to some linear function of the expected value of the vector of observations, $\boldsymbol{y}$."

Thus a scalar parameter, $\theta_i = \mathbf{C}_{1\times p}\, \boldsymbol{\beta}_{p\times 1}$, is estimable

$$\Leftrightarrow \mathbf{C}_{1\times p}\, \boldsymbol{\beta}_{p\times 1} = \mathbf{t}'_{1\times n} E(\mathbf{y}_{n\times 1}),$$

for $\boldsymbol{t}$ a vector of constants.

More generally, for a vector we need $\boldsymbol{\theta}_{\boldsymbol{a}\times \mathbf{1}} = \boldsymbol{T}_{\boldsymbol{a}\times \boldsymbol{n}} E\left(\boldsymbol{y}_{n\times 1}\right)$

There always exist $r =$rank$(\boldsymbol{X})$ distinct and estimable parameters (which are not necessarily elements of $\boldsymbol{\beta}$ but may be linear combinations of elements).

If rank$(\boldsymbol{X}) = r = p$ , then $\widehat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable, and any (nonzero) $\boldsymbol{C}$ gives estimable $\boldsymbol{\theta}$. This is usually the case with continuous predictors unless some predictors are collinear.

If rank$(\boldsymbol{X}) = r < p$, $\boldsymbol{\beta}$ is not estimable (although as many as $r$ elements may be), and for $\widehat{\boldsymbol{\theta}} = \boldsymbol{C}\boldsymbol{\beta}$, we must check estimability.

To show set of parameters

$$\boldsymbol{\theta}_{a\times 1} = \mathbf{C}_{a\times p}\, \boldsymbol{\beta}_{p\times 1} = \mathbf{T}_{a\times n} E(\mathbf{y}_{n\times 1})$$

is estimable, it suffices to show that $\mathbf{C}_{a\times p} = \mathbf{T}_{a\times n}\mathbf{X}_{n\times p}$.

Estimable $\widehat{\boldsymbol{\theta}}$ shares the optimality of $\widehat{\boldsymbol{\beta}}$ (whatever $r$ is): BLUE for least squares and MLE with Gaussian errors.

Show that $H_0 : \beta_1 = \beta_2$ is estimable for the ozone data.

## Testability of a Hypothesis

Consider the likelihood ratio (LR) test. Let $\mathbf{M}_{a\times a} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$.
Define GLH testability as the (unique) existence of the LR test.

$\boldsymbol{\theta}$ is testable $\Leftrightarrow$

- $\boldsymbol{C}$ is full rank $a$ (no redundancies), and
- $\boldsymbol{\theta}$ is estimable,

OR, equivalently,

- $\boldsymbol{M}$ is full rank $a$, and
- $\boldsymbol{\theta}$ is estimable.

If $\boldsymbol{X}$ is full rank, then $\boldsymbol{\theta}$ is testable $\Leftrightarrow$
$\boldsymbol{C}$ is full rank $a$ OR $\boldsymbol{M}$ is full rank $a$ (because any $\boldsymbol{\theta}$ is estimable)

Show that $H_0 : \beta_1 = \beta_2$ is testable for the ozone data.

## Computation of Test Statistic and p-Value

Define the sums of squares hypothesis as
$SSH_{1\times 1} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

With HILE Gauss, the likelihood ratio statistic equals

$$\begin{aligned} F_{obs} &= \frac{SSH/a}{SSE/(n-r)} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\widehat{\sigma}^2} \\ &= \frac{MSH}{MSE} \end{aligned}$$

Under $H_0$: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $SSH$ and $SSE$ are scaled $\chi^2$ random variables, with $SSH/\sigma^2 \sim \chi^2(a)$, independently of $SSE/\sigma^2 \sim \chi^2(n-r)$. It can be shown that if $z_1 \sim \chi^2_{d_1}$, $z_2 \sim \chi^2_{d_2}$, and $z_1 \perp z_2$, then $\frac{z_1/d_1}{z_2/d_2}$ follows an $F_{d1,d2}$ distribution. Thus

$$F_{obs} = \frac{[SSH/\sigma^2]/a}{[SSE/\sigma^2]/(n-r)} = \frac{SSH/a}{SSE/(n-r)} \sim F(a, n-r).$$

The p-value equals the probability of observed or more extreme data arising under the null, that is,

$$\text{p-val} = \Pr\{F(a, n-r) \geq F_{obs}\} = 1 - \Pr\{F(a, n-r) < F_{obs}\}.$$

Reject $H_0$ if $F_{obs} > f_{crit} = F^{-1}(1-\alpha, a, n-r)$.

Obtain $f_{crit}$ in SAS as FINV(prob, $df_1, df_2$), the value of an $F$ statistic with $df_1$ numerator and $df_2$ denominator degrees of freedom, such that $\Pr\{F \leq f_{crit}\} =$prob. In R, use qf(prob, $df_1$, $df_2$). (To get p-values in R, use 1 - pf(crit, $df_1$, $df_2$).

All linear model GLH tests correspond to comparing two models, the "full" model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and a reduced model defined by constraints. This concept lies at the heart of any LR test and is critical in understanding any particular GLH test.

## Example: Computing a GLH Test

For the ozone data, again consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $O_{PERSONAL}$ as the response and $O_{OUTDOOR}$, $O_{HOME}$, and $TIME_{OUTDOORS}$ as predictors. Suppose we wish to test that all slopes are equal. Then $H_0 : \boldsymbol{\theta} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix} = \mathbf{0}$, with corresponding contrast matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$.

The additional code needed to fit the model, along with PROC REG code, is given below.

```
> C = matrix(c(0,1,-1,0,0,0,1,-1), nrow = 2, byrow = T) # contrast matrix
> print(C)

     [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1

> M=C %*% solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(thetahat)%*%solve(M)%*%thetahat
> f_obs=(ssh/nrow(thetahat))/mse
> p=1-pf(f_obs,2,60)
```

The results from R are below.

```
> print(ssh)

         [,1]
[1,] 1237.324

> print(f_obs)

         [,1]
[1,] 3.657767

> print(p)

           [,1]
[1,] 0.03170141
```

We reject the null hypothesis and conclude that not all slopes are identical. At least one slope is not equal to the others.

This output corresponds to PROC REG below.

```
proc reg data=ozone;
model personal=outdoor home time_out;
test outdoor-home=0, home-time_out=0;
run;
```

```
                    The REG Procedure
                      Model: MODEL1

        Test 1 Results for Dependent Variable personal

                                  Mean
Source              DF           Square    F Value    Pr > F

Numerator            2        618.66202       3.66    0.0317
Denominator         60        169.13652
```

## Wald Tests

For a single coefficient $\beta_j$, we can test $H_0 : \beta_j = 0$ if $\beta_j$ is estimable. SAS automatically reports *Wald* tests for parameters in a regression model. These are tests of the hypothesis $H_0 : \beta_j = 0$ for each $j$.

We know that in large samples (or in small samples if our errors are exactly normal), $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. The variance of $\widehat{\beta}_j$ is the $j^{th}$ diagonal element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Using properties of the standard normal distribution, we can base our test on the ratio

$$t = \frac{\widehat{\beta}_j - 0}{\sqrt{\mathsf{var}(\widehat{\beta}_j)}}.$$

Usually, we do not know $\sigma^2$ exactly and obtain an estimate as $\widehat{\sigma}^2 = \frac{SSE}{dfE}$. If we do know $\sigma^2$ exactly, then $t \sim N(0,1)$. If we estimate $\sigma^2$ from the data, then $t \sim t_{dfE}$, a Student's $t$ distribution with $dfE$ degrees of freedom.

## One and Two-Sided Tests

One-sided tests exist only for scalar hypotheses, not for vector hypotheses. Let $a =\#$ rows of $\boldsymbol{C}$. If $a = 1$, then $\boldsymbol{\theta}$ is a scalar $(1 \times 1)$ and $F_{obs}(1, n-r) = t^2(n-r)$, where $t(n-r)$ is t-statistic with d.f. of $n-r$.

- Two-sided test: $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$
- One-sided tests: $\begin{cases} H_0\text{:} \quad \theta = \theta_0 \quad \text{vs.} \quad H_\text{A}\text{:} \quad \theta > \theta_0 \\ H_0\text{:} \quad \theta = \theta_0 \quad \text{vs.} \quad H_\text{A}\text{:} \quad \theta < \theta_0 \end{cases}$

*Conducting a test of size $\alpha$ by t-test:*

- A two-sided t-test, $H_\text{A}\text{:} \quad \theta \neq \theta_0$, uses the $\alpha/2$ and $(1-\alpha/2)$ critical values.
- A one-sided t-test,, $H_\text{A}\text{:} \quad \theta < \theta_0$, uses the $\alpha$ critical value.
- A one-sided t-test,, $H_\text{A}\text{:} \quad \theta > \theta_0$, uses the $(1-\alpha)$ critical value.

In all cases, one rejects $H_0$ if the test statistic is farther from zero than

the critical value.

F-tests

- A two-sided $F$ test, $H_{\mathsf{A}}$: $\theta \neq \theta_0$, uses the $\alpha$ critical value.
- A one-sided $F$ test, $H_{\mathsf{A}}$: $\theta < \theta_0$, uses the $2\alpha$ critical value and also requires appropriate sign of the difference.

Recall the definition of a CDF: $F_X(x) = \Pr\{X \leq x\}$.

$$F_F(f) = \Pr\{F \leq f\} = \int_0^f f_F(u)du$$

2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

## Next: Distributional Results

*Reading Assignment:*

- Muller and Fetterman Chapter 3: "Some Distributions for the GLM" (Required)

# Lecture 6: Some Distributional Results for the GLM

*Reading Assignment:*

- Muller and Fetterman Chapter 3: "Some Distributions for the GLM"

In analysis with the GLM, we use three kinds of distributions: multivariate Gaussian, $\chi^2$, and $F$.

For now, assume all assumptions hold:

- HILE for estimation and
- Gaussian errors for testing.

## A Full Rank Basis for Less Than Full Rank Models

If $\boldsymbol{X}$ is $n \times p$, with $n \geq p$ and rank$(\boldsymbol{X}) = r \leq p$, then rank$(\boldsymbol{X}'\boldsymbol{X}) = r$.

If $\boldsymbol{X}$ is less than full rank ($r < p$), then *collinearity* exists among columns of $\boldsymbol{X}$. If $\mathbf{X}$ is less than full rank, then we say the model is also less than full rank. Also, $r =$ rank$(\boldsymbol{X}) =$ rank$(\mathbf{X}'\mathbf{X})$ is the # of estimable parameters.

For every less than full rank model, there exists a corresponding full rank model with $r$ estimable parameters. That is, for less than full rank $\boldsymbol{X}$, there exists a $p \times r$ matrix $\boldsymbol{V}_+$ such that

$$\mathbf{X}_{n\times p} = \boldsymbol{X}_{*,(n\times r)}\mathbf{V}'_{+,(r\times p)}$$

with rank$(\boldsymbol{X}_*)$ =rank$(\boldsymbol{V}_+) = r < p$. $\boldsymbol{X}_*$ provides a *full rank basis* for $\boldsymbol{X}$.

Suppose that we have the model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\ \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

where rank$(\mathbf{X}) = r < p$.

Then, defining $\mathbf{X}_{*,(n\times r)} = \mathbf{X}_{n\times p}\ \mathbf{V}_{+,(p\times r)}$ with corresponding parameter vector $\boldsymbol{\beta}_{*,(r\times 1)}$, an equivalent full-rank model is given by

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{*,(n\times r)}\ \boldsymbol{\beta}_{*,(r\times 1)} + \boldsymbol{\varepsilon}_{n\times 1},$$

with $\widehat{\boldsymbol{\beta}_*} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{y}$.

Many possible choices of the matrix $\boldsymbol{V}_+$ exist, such as the set of eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$ corresponding to non-zero eigenvalues.

Every parameter estimable in the original (less than full rank) model is also estimable in the full rank model, and any estimable parameter is expressible as a linear combination of the $\beta_*$'s.

## More About the Multivariate Gaussian Distribution

If $\mathbf{\Sigma}$ is not full rank (e.g.,covariance matrix of residuals), then the multivariate Gaussian distribution is said to be *singular normal*. In this case, the density does not exist. (The multivariate Gaussian density exists only when $\mathbf{\Sigma}_z$ is non-singular. For example, if $y \sim N_1(\mu, 0)$, we have a discrete distribution with $P(y = \mu) = 1$, a point mass at $\mu$.)

We define a singular multivariate Gaussian distribution for a vector $\mathbf{z}$ in terms of a particular linear transformation $\mathbf{Uz}$ that leads to a full rank covariance matrix $\mathbf{U\Sigma U}'$ so that we can define a density for the transformed random vector $\mathbf{Uz}$ (with redundancies eliminated).

## Definition of Singular and Nonsingular Multivariate Gaussian

1. Full Rank Case

   If rank$(\boldsymbol{\Sigma}_z) = n$ and the density of $\boldsymbol{z}$ is

   $$p(\boldsymbol{z}) = (2\pi)^{-n/2} \left|\boldsymbol{\Sigma}_z\right|^{-1/2} \exp[-(\boldsymbol{z} - \boldsymbol{\mu}_z)'\boldsymbol{\Sigma}_z^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_z)/2],$$

   then $\boldsymbol{z}$ is distributed multivariate Gaussian, indicated

   $$\mathbf{z}_{n\times 1} \sim \mathcal{N}_n(\boldsymbol{\mu}_{z,(n\times 1)}, \boldsymbol{\Sigma}_{z,(n\times n)}).$$

   For example, with $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and HILE Gauss, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$ and $\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$.

2. Less than Full Rank Case

   If $1 \leq$ rank$(\boldsymbol{\Sigma}_z) = n_* < n$, then a fixed matrix $\boldsymbol{U}_{n_*\times n}$ exists such that rank$\big(\boldsymbol{U}\boldsymbol{\Sigma}_z\boldsymbol{U}'\big) = n_*$ (full rank).

If the density of $\boldsymbol{U}\boldsymbol{z}$ is

$$\begin{aligned} p(\mathbf{U}\mathbf{z}) &= (2\pi)^{-\frac{n_*}{2}} \mid \mathbf{U}\boldsymbol{\Sigma}_z\mathbf{U}' \mid^{-\frac{1}{2}} \\ &\quad \exp\left[-\frac{1}{2}(\mathbf{U}\mathbf{z} - \mathbf{U}\boldsymbol{\mu}_z)'(\mathbf{U}\boldsymbol{\Sigma}_z\mathbf{U}')^{-1}(\mathbf{U}\mathbf{z} - \mathbf{U}\boldsymbol{\mu}_z)\right], \end{aligned}$$

then $\mathbf{z}$ is distributed as a singular multivariate Gaussian, indicated by

$$\mathbf{z}_{n\times 1} \sim \mathcal{SN}_n(\boldsymbol{\mu}_{z,(n\times 1)}, \boldsymbol{\Sigma}_{z,(n\times n)}).$$

We must add side conditions to specify $\boldsymbol{U}_{n_*\times n}$ uniquely.

Example: Pick $\boldsymbol{U}$ to contain the eigenvectors for non-zero eigenvalues of $\boldsymbol{\Sigma}_z$.

## Sampling Distributions of Estimators

$\boldsymbol{\beta}$ **Estimators** $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

From properties of the multivariate Gaussian distribution,

$$\begin{aligned} E[\widehat{\boldsymbol{\beta}}] &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] E(\mathbf{y}) \\ &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}, \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \mathrm{Cov}(\mathbf{y}) \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\ &= \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] (\sigma^2\mathbf{I}_n) \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

with

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

because $\widehat{\boldsymbol{\beta}}$ is a full rank linear transformation of the normal random vector $\mathbf{y}$.

For $\mathbf{X}$ less than full rank and appropriate choice of generalized inverse, we have

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{SN}_p \left[ (\mathbf{X}'\mathbf{X})^{-}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-} \right],$$

which is a singular multivariate Gaussian because $(\boldsymbol{X}'\boldsymbol{X})$ is less than full rank.

For a Gaussian distribution, we know that most ($> 95\%$) of its mass lies within two standard deviations of the mean. So if our null hypothesis is for no effect of the covariate corresponding to $\beta_j$ and

$$|\widehat{\beta}_j| > 0 + 1.96 * \sqrt{\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}_{j,j}},$$

where $(\mathbf{X}'\mathbf{X})^{-1}_{j,j}$ refers to the $(j,j)^{th}$ element of $(\mathbf{X}'\mathbf{X})^{-1}$, then we have evidence that the covariate of interest has an effect on the response.

**Example: Covariance Matrix of $\widehat{\boldsymbol{\beta}}$**

The following R and PROC REG code below may be used to obtain the estimated covariance matrix of $\widehat{\boldsymbol{\beta}}$ for the ozone data with $O_{PERSONAL}$ as the outcome and $O_{OUTDOOR}$, $O_{HOME}$, and $TIME_{OUTDOORS}$ as predictors.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)
  (Intercept)  outdoor  home time_out
1           1 35.87771 22.29     0.57
2           1 43.79189 13.97     0.90
3           1 49.81255 18.96     0.55
4           1 34.37366 22.27     0.17
5           1 45.95496 23.40     0.00
6           1 64.76558 39.62     0.30
> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
> covbhat=solve(t(X) %*% X)*as.numeric(mse) # scalar multiplication
> print(covbhat)
```

```
            (Intercept)      outdoor        home     time_out
(Intercept)  18.8534457 -0.183111754 -0.18665969 -15.18741728
outdoor      -0.1831118  0.008175203 -0.00831528  -0.06888717
home         -0.1866597 -0.008315280  0.02715354   0.07755804
time_out    -15.1874173 -0.068887173  0.07755804  59.43995287
```

We may also obtain the estimated covariance matrix of the $\beta$'s using the COVB option in SAS PROC REG.

```
proc reg;
model personal=outdoor home time_out/covb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: personal

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 5034.90667 | 1678.30222 | 9.92 | <.0001 |
| Error | 60 | 10148 | 169.13652 | | |
| Corrected Total | 63 | 15183 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 13.00525 | R-Square | 0.3316 |
| Dependent Mean | 23.54578 | Adj R-Sq | 0.2982 |
| Coeff Var | 55.23389 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 3.78349 | 4.34206 | 0.87 | 0.3870 |
| outdoor | 1 | 0.09142 | 0.09042 | 1.01 | 0.3160 |
| home | 1 | 0.59544 | 0.16478 | 3.61 | 0.0006 |

| | | | | | |
|---|---|---|---|---|---|
| time_out | 1 | 13.64454 | 7.70973 | 1.77 | 0.0818 |

Model: MODEL1
Dependent Variable: personal Personal Ozone Exposure (ppb)

Covariance of Estimates

| Variable | Label | Intercept | outdoor |
|---|---|---|---|
| Intercept | Intercept | 18.853445739 | -0.183111754 |
| outdoor | Outdoor Ozone Concentration (ppb) | -0.183111754 | 0.0081752029 |
| home | Home Indoor Ozone Concentration (ppb) | -0.186659692 | -0.00831528 |
| time_out | Proportion of Time Spent Outdoors | -15.18741728 | -0.068887173 |

Covariance of Estimates

| Variable | Label | home | time_out |
|---|---|---|---|
| Intercept | Intercept | -0.186659692 | -15.18741728 |
| outdoor | Outdoor Ozone Concentration (ppb) | -0.00831528 | -0.068887173 |
| home | Home Indoor Ozone Concentration (ppb) | 0.0271535425 | 0.0775580389 |
| time_out | Proportion of Time Spent Outdoors | 0.0775580389 | 59.439952865 |

## $\boldsymbol{\theta}$ Estimator

Again using properties of the multivariate Gaussian distribution, for $\mathbf{C}_{a\times p}$ a matrix of constants and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\theta}} \sim N_a(\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}').$$

We should be careful to verify that $\boldsymbol{\theta}$ is both estimable and testable.

## Example: Estimating $\boldsymbol{\theta}$

Suppose we wish to test the hypothesis that the $O_{OUTDOOR}$, $O_{HOME}$, and $TIME_{OUTDOORS}$ coefficients are all equal in the ozone data. So

$$\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix},$$

and

$$\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_2(\boldsymbol{\theta}, \sigma^2 \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}').$$

We may obtain the estimated covariance matrix of $\widehat{\boldsymbol{\theta}}$ with the following additional R code.

```
> C=matrix(c(0, 1, -1, 0, 0, 0, 1, -1), nrow = 2, byrow = T);
> print(C)

     [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1

> covhat=as.numeric(mse)*C%*%solve(t(X) %*% X)%*%t(C) # sigma^2 * M
> print(covhat)

           [,1]       [,2]
[1,] 0.05195931  0.1109764
[2,] 0.11097639 59.3119903
```

## Predicted Values: Conditional Means and Future Observations

For the GLM

$$\begin{aligned} E(\mathbf{y}_{n\times 1}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}(\mathbf{y} \mid \mathbf{X}). \end{aligned}$$

We write the estimator of the expected values as

$$\begin{aligned} \widehat{\mathbf{y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} \\ &= \mathbf{H}\mathbf{y}. \end{aligned}$$

Prediction is often an important goal of modeling. For example, an actuary might want to predict medical costs for health insurance patients covered by a given insurance plan given their characteristics (including BMI, smoking status, and age).

Caution: prediction outside the range of observed data can be extremely dangerous!

To find the distribution of $\widehat{\mathbf{y}}$,

$$\begin{aligned} E(\widehat{\mathbf{y}}) &= \mathbf{H}E(\mathbf{y}) \\ &= \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

In addition, for $\widehat{\mathbf{y}}$ an estimated conditional mean response,

$$\begin{aligned} \text{cov}(\widehat{\mathbf{y}}) &= \mathbf{H}(\sigma^2\mathbf{I})\mathbf{H}' \\ &= \sigma^2\mathbf{H}\mathbf{H}' = \sigma^2\mathbf{H} \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \end{aligned}$$

as $\mathbf{H}$ is symmetric and idempotent.

As a transformation of a singular multivariate Gaussian ($\mathbf{H}_{n\times n}$ has rank $p < n$ when $\mathbf{X}$ is full rank), the distribution of $\widehat{\mathbf{y}}$ is given by

$$\widehat{\mathbf{y}} \sim SN_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}).$$

In contrast, consider predicting future observations at time $f$ with covariate values $\mathbf{X}_f$. In this case, we predict an *individual outcome* rather than a mean outcome. The predicted future response $\widehat{\boldsymbol{y}}_f$ involves variance due to

1. estimating $\boldsymbol{\beta}$ (the fitted line is not the exact true line)
2. observing $\boldsymbol{\varepsilon}_f$ in the future sample (even if the fitted line is the exact true line, there is still variability about it), and
3. changing the design (predictor values).

Now

$$\widehat{\boldsymbol{y}}_f = \boldsymbol{X}_f \widehat{\boldsymbol{\beta}},$$

where $\boldsymbol{X}_f$ is $n_f \times p$, while $\boldsymbol{X}$ is $n \times p$. ($\boldsymbol{X}_f \neq \boldsymbol{X}$). These predictions will have additional error (as opposed to estimated conditional means) $\boldsymbol{\varepsilon}_f$: the errors to be observed at future time $f$ ($\boldsymbol{\varepsilon}_f \neq \boldsymbol{\varepsilon}$), and $\boldsymbol{\varepsilon}_f$ are independent of the errors at the current time.

Thus we have

$$\widehat{\boldsymbol{y}}_f \sim \mathcal{SN}_{n_f}\{\boldsymbol{X}_f\boldsymbol{\beta},\, \sigma^2[\boldsymbol{X}_f(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_f' + \boldsymbol{I}]\}$$

$\boldsymbol{X}_f = \boldsymbol{X}$ implies

$$\widehat{\boldsymbol{y}}_f \sim \mathcal{SN}_n\{\boldsymbol{X}\boldsymbol{\beta},\, \sigma^2[\boldsymbol{H} + \boldsymbol{I}]\},$$

so we see the additional uncertainty in predicting a particular future observation rather than estimating the mean.

It would be unambiguous to discuss estimated conditional means, $\widehat{\boldsymbol{\mu}}\big|\boldsymbol{X}$, or predicted future observations, $\widehat{\boldsymbol{y}}_f$; however, describing $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ as predicted values has become standard, so that you must rely on the context to decide whether estimated conditional means or predicted future observations are being discussed.

## Definitions and Properties of Residuals

The residuals $\widehat{\boldsymbol{\varepsilon}}$ are defined as

$$\begin{aligned}\widehat{\boldsymbol{\varepsilon}} &= (\mathbf{y} - \widehat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y},\end{aligned}$$

with

$$\widehat{\boldsymbol{\varepsilon}} \sim \mathcal{SN}_n(\mathbf{0}, \sigma^{\mathbf{2}}(\mathbf{I} - \mathbf{H})).$$

$\widehat{\varepsilon}_i$ is not independent of $\widehat{\varepsilon}_j$ $(i \neq j)$ unless $p = 0$ or $n \to \infty$.

$\widehat{\varepsilon}_i \sim \mathcal{N}[0, \sigma^2(1 - h_i)]$, where $h_i = \mathbf{x}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\mathbf{x}_i'$, the $(i, i)$ element of $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

If the model spans an intercept, $\sum_{i=1}^n \widehat{\varepsilon}_i = 0$.

## Residual Variance

Let $r = \text{rank}(\mathbf{X})$. Recall that

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{SSE}{n-r} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-r} \\
&= \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-r} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-r}
\end{aligned}$$

as $(\mathbf{I}-\mathbf{H})$ is symmetric and idempotent.

Now

$$\widehat{\sigma}^2 \left(\frac{n-r}{\sigma^2}\right) \sim \chi^2(n-r).$$

Because $E(\chi^2(\nu)) = \nu$, $\widehat{\sigma}^2$ is unbiased. In addition, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are statistically independent.

## Standardized Residuals

Replacing $\sigma^2$ with $\widehat{\sigma}^2$ leads some distributions to change from Gaussian to Student's $T$.

Because $\widehat{\varepsilon}_i \sim \mathcal{N}[0, \sigma^2(1-h_i)]$, it follows that

$$\frac{\widehat{\varepsilon}_i}{\sqrt{\sigma^2(1-h_i)}} \sim \mathcal{N}(0,1).$$

Define the standardized residual as

$$r_i = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2(1-h_i)}}$$

The standardized residuals do not follow a $T$ distribution; in fact,

$$\frac{r_i^2}{n-r} \sim \text{Beta}\left(\frac{1}{2}, \frac{n-r-1}{2}\right),$$

and the $r_i$ are bounded between $-\sqrt{n-r}$ and $\sqrt{n-r}$, where $r = \text{rank}(\mathbf{X})$.

## Example: Calculating Standardized Residuals

To calculate the standardized residuals for the ozone example, we add the following R code to the previous code. In addition, we may obtain the standardized residuals from PROC REG by using the "R" option with the model statement.

```
> H=X%*%solve(t(X) %*% X) %*% t(X) # calculate Hat Matrix
> h_i=diag(H) # get the diagonal
> r_i=ehat/(sqrt(as.numeric(mse)*(1-h_i)));
> head(r_i,5)
```

```
        [,1]
1 -0.1441837
2 -1.1486537
3 -0.4461650
4 -1.4976573
5  0.5791377
```

The first five residuals are printed above. You may also get these in R from your fitted linear model using the MASS package.

```
> out = lm(personal ~ outdoor + home + time_out, data = ozone)
> # equivalent to out = lm(y ~ X - 1)
```

```
> summary(out)

Call:
lm(formula = personal ~ outdoor + home + time_out, data = ozone)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.930  -7.855  -4.257   4.880  36.295 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.78349    4.34206   0.871 0.387032    
outdoor      0.09142    0.09042   1.011 0.316031    
home         0.59544    0.16478   3.613 0.000619 ***
time_out    13.64454    7.70973   1.770 0.081845 .  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 13.01 on 60 degrees of freedom
Multiple R-squared:  0.3316,	Adjusted R-squared:  0.2982 
F-statistic: 9.923 on 3 and 60 DF,  p-value: 2.094e-05

> library(MASS) #load MASS package, or install using install.packages("MASS")
```

```
> r_i_MASS = stdres(out) # stdres function from MASS
> head(r_i_MASS,5)

         1          2          3          4          5
-0.1441837 -1.1486537 -0.4461650 -1.4976573  0.5791377
```

The following PROC REG code obtain the residuals.

```
proc reg;
model personal=outdoor home time_out/r;
run;
```

The first five residuals are below. Note that the standardized residuals are labeled 'student residual.'

```
The REG Procedure
                                Model: MODEL1
          Dependent Variable: personal Personal Ozone Exposure (ppb)

                              Output Statistics

              Dep Var  Predicted     Std Error              Std Error   Student
       Obs   personal      Value  Mean Predict   Residual   Residual  Residual

         1    26.2900    28.1131        3.0429    -1.8231     12.644    -0.144
         4     3.3000    22.5059        2.1640   -19.2059     12.824    -1.498
         5    29.2800    21.9179        2.7454     7.3621     12.712     0.579
         9    28.5500    20.0446        5.9850     8.5054     11.546     0.737
        13    38.2800    37.0471        5.1369     1.2329     11.948     0.103
```

## Jackknifing

Jackknifing usually involves computing a statistic with one observation deleted from a sample, once for each observation. Let the subscript $(-i)$ indicate having deleted the $i$th observation. For example, $\boldsymbol{y}_{(-i)}$ and $\boldsymbol{X}_{(-i)}$ have $n-1$ rows.

Compute

- $\widehat{\boldsymbol{\beta}}_{(-i)} = (\boldsymbol{X}'_{(-i)}\boldsymbol{X}_{(-i)})^{-1}\boldsymbol{X}'_{(-i)}\boldsymbol{y}_{(-i)}$
- $\widehat{\boldsymbol{y}}_{(-i)} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(-i)}$
- $\widehat{\boldsymbol{\varepsilon}}_{(-i)} = \boldsymbol{y}_{(-i)} - \widehat{\boldsymbol{y}}_{(-i)}$ and
- $\widehat{\sigma}^2_{(-i)} = \widehat{\boldsymbol{\varepsilon}}'_{(-i)}\widehat{\boldsymbol{\varepsilon}}_{(-i)}/(n-r-1)$,

where $r = \text{rank}(\mathbf{X})$.

If $\widehat{\boldsymbol{\beta}}_{(-i)}$ differs greatly from $\widehat{\boldsymbol{\beta}}$, then we see that observation $i$ has a good deal of influence on the analysis.

## Studentized Residuals

Standardizing the jackknifed residuals (by dividing by an estimate of the standard deviation) yields a set of residuals (the *Studentized residuals*) which follow a Student's $T$ distribution:

$$r_{(-i)} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2_{(-i)}(1-h_i)}} \sim T(n-r-1)$$

Chapter 7 in MF has extensive discussion of this most important tool for assumption evaluation.

## Example: Calculating Studentized Residuals

The following code may be added to the previous code to obtain the studentized residuals in PROC IML. The code uses the following formula for the studentized residuals given in Muller and Fetterman:

$$r_{(-i)} = r_i \left( \frac{(n-r)-1}{(n-r)-r_i^2} \right)^{\frac{1}{2}},$$

where $r = \text{rank}(\mathbf{X})$.

```
> library(Matrix) #install.packages("Matrix")
> r_i2=r_i^2; # square each element of r_i
> r = rankMatrix(X) # full rank
> r_mi=r_i*sqrt((n-r-1)/(n-r-r_i2))
> head(r_mi,5)
```

```
        [,1]
1 -0.1430019
2 -1.1517756
3 -0.4431671
4 -1.5136869
5  0.5759032
```

The following SAS PROC REG code may be used to obtain the studentized residuals.

```
proc reg;
model personal=outdoor home time_out;
output out=resid student=standresid rstudent=studresid;
run;

proc print data=resid;
var personal standresid studresid;
run;
```

The first five observations are given below.

| Obs | personal | standresid | studresid |
|---|---|---|---|
| 1 | 26.29 | -0.14418 | -0.14300 |
| 4 | 3.30 | -1.49766 | -1.51369 |
| 5 | 29.28 | 0.57914 | 0.57590 |
| 9 | 28.55 | 0.73663 | 0.73379 |
| 13 | 38.28 | 0.10319 | 0.10233 |

## Next: Multiple Regression

*Reading Assignment:*

- Muller and Fetterman, Chapter 4: "Multiple Regression"

# Lecture 7: Multiple Regression: General Consideration

*Reading Assignment:*

- Muller and Fetterman, Chapter 4: "Multiple Regression"

Why use more than one covariate in a model?

- Why not fit separate models for every covariate?
- Omitting an important covariate, $\mathbf{x}_2$, can cause you to miss significant relationships between $\mathbf{x}_1$ and $\mathbf{y}$ or even to make completely wrong conclusions!

## Example: Math Ability

Suppose that you hypothesize that taller children are better at math than shorter children. You take a random sample of 32 children of various ages in Ephesus Elementary. These children take a math test and have their heights measured. When you fit a linear model using height to predict math test score, you find that height is highly significant.

Does this make sense? Recall the definition of a confounder: a *confounder* is a factor that is associated with the exposure and independently affects the outcome.

In this case, age is a potential confounder because it is associated with height and, independently of height, is related to math test scores.

When we fit the *multiple regression* model (sometimes called a *multivariable* regression model) with both age and height as predictors, we see that age is an important predictor of mathematical ability. In addition, after accounting for age, height is unimportant. Our conclusion is that after accounting for age, height is not related to mathematical ability in our sample of elementary school children.

For a general multiple regression model, we write

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & \cdots & x_{1,p-1} \\ \vdots & & \\ x_{n0} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The corresponding scalar version of the model is

$$y_i = \sum_{j=0}^{p-1} x_{ij}\beta_j + \varepsilon_i$$

Assume we fit a model and test hypotheses with a continuous, interval scale response and with linear combinations of one or more continuous variables as predictors. All GLH tests can be understood in terms of comparisons of two models: a full model and a reduced model. These tests may be conducted by comparing the *sums of squares* from the two models.

## Definitions of Basic Sums of Squares

For the model $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, we have

- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,
- $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, and
- $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$.

We are already familiar with the sum of squares for error, given by

$$\begin{aligned} SSE &= \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \widehat{\mathbf{y}})'(\mathbf{y} - \widehat{\mathbf{y}}) \\ &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}\widehat{\varepsilon}_i^2. \end{aligned}$$

The *uncorrected total sum of squares*, $USS$, is given by

$$USS(\text{total}) = \boldsymbol{y}'\boldsymbol{y} = \sum_{i=1}^{n} y_i^2.$$

The *uncorrected model sum of squares* (or *uncorrected regression sum of squares*) is given by

$$USS(\text{model}) = \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} = \sum_{i=1}^{n} \widehat{y}_i^2.$$

The difference between the uncorrected total sum of squares and the uncorrected model sum of squares is the sum of squares for error:

$$USS(\text{total}) - USS(\text{model}) = \mathbf{y}'\mathbf{I}\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = SSE.$$

So we see that

$$\begin{aligned} USS(\text{total}) &= USS(\text{model}) + SSE \\ \mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

## Example: Calculating Uncorrected Sums of Squares

Uncorrected sums of squares may be calculated using the following R code.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)
```

```
  (Intercept)  outdoor  home time_out
1           1 35.87771 22.29     0.57
2           1 43.79189 13.97     0.90
3           1 49.81255 18.96     0.55
4           1 34.37366 22.27     0.17
5           1 45.95496 23.40     0.00
6           1 64.76558 39.62     0.30
```

```
> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
```

```
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
> H=X%*%solve(t(X) %*% X) %*% t(X) # calculate Hat Matrix
> uss_t=t(y)%*%y
> uss_m=t(y)%*%H%*%y
> print(uss_t)

         [,1]
[1,] 50664.94

> print(uss_m);

         [,1]
[1,] 40516.75

> print(uss_m + sse);

         [,1]
[1,] 50664.94
```

We may also get the same results in SAS using PROC GLM, a more general form of PROC REG, using the int option to calculate the uncorrected sums of squares, printing them in an ANOVA table.

```
proc glm data=ozone;
model personal=outdoor home time_out/int;
run;
```

The GLM Procedure

Dependent Variable: personal  Personal Ozone Exposure (ppb)

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 4 | 40516.75081 | 10129.18770 | 59.89 | <.0001 |
| Error | 60 | 10148.19129 | 169.13652 | | |
| Uncorrected Total | 64 | 50664.94210 | | | |

Corrected sums of squares are defined in relationship to the intercept.

## The Nature of the Intercept

It is often (but not always) appropriate to include a constant predictor equal to 1.0 so that $x_0 = \mathbf{J}$. Then

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{i,p-1}\beta_{p-1} + \varepsilon_i.$$

In such models, we call $\beta_0$ the intercept. The name reflects the fact that if $x_{i0} = 1$ and all other predictors take the value zero, then

$$E(y_i) \quad = \quad E(\beta_0) + E\left(\sum_{j=1}^{p-1} x_{ij}\beta_j\right) + E(\varepsilon_i) = \beta_0,$$

and $\beta_0$ is the $y$ intercept of the fitted line (the value of $y$ at which the regression function intercepts the vertical axis). In addition, $\beta_0$ equals the response ($y$) value predicted if all $x_{ij}$, $j \in \{1, 2, \ldots p-1\}$, are 0.

For $\bar{x}_j$ the average of the $j^{th}$ column of $\mathbf{X}$, $\beta_0 = \mu_y - \sum_{j=1}^{p-1} \overline{x}_j \beta_j$ and $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p-1} \overline{x}_j \widehat{\beta}_j$.

## Example: Ozone Data

For the ozone data, we have the model

$$E(personal_i) = \beta_0 + \beta_1 outdoor_i + \beta_2 home_i + \beta_3 timeout_i,$$

$i = 1, \ldots, 64$. The parameter estimates from this model are given by $\widehat{\boldsymbol{\beta}} = (3.78, 0.09, 0.60, 13.64)'$. In addition, in our dataset, we observe outdoor ozone concentrations ranging from 11.60-104.10 ppb, home ozone concentrations ranging from 0.80-46.04 ppb, and percent time spent outdoors ranging from 0-90%. In this model, are we particularly interested in the value of the intercept?

## Models That Span but May Not Include an Intercept

Any model that includes an intercept has $\mathbf{J}_n$ as a column in $\boldsymbol{X}$. Models with an intercept allow computing the corrected sums of squares, which exclude the portion of the sums of squares due to the intercept and are usually preferred over the uncorrected sums of squares.

When using dummy variables, it is often convenient to code the model such that no column of $\boldsymbol{X}$ is $\mathbf{1}$ although $\mathbf{X}_{n\times p}\ \mathbf{t}_{p\times 1} = \mathbf{1}$, for $\mathbf{t}_{p\times 1}$ a vector of constants.

Such a model *spans* an intercept, even though the design matrix, $\boldsymbol{X}$, does not include an intercept.

For example, let

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{\beta}_1 = \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \ \boldsymbol{X}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \ \boldsymbol{\beta}_2 = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}.$$

Notice that $\mathbf{X}_2 = \mathbf{X}_1 \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_0 + \alpha_1 \end{bmatrix}$$

so that the vector $\mathbf{t} = (1, 1)'$ finds the hidden intercept.

## Corrected Sums of Squares

Corrected sums of squares are adjusted (corrected) for the intercept and measure the effect above and beyond it. In addition, the corrected sums of squares are invariant to location shifts of the response or predictors (e.g.,centering a covariate).

Uncorrected sums of squares are always well-defined. Corrected sums of squares and associated statistics entirely exclude the intercept but involve comparing models which all span an intercept. Thus corrected sums of squares are defined only if the model includes or spans an intercept.

The correction term or *sum of squares due to the intercept*, $n\overline{y}^2 = \frac{\boldsymbol{y}'\mathbf{J}_n\mathbf{J}'_n\boldsymbol{y}}{n} = SSI$, corrects for location.

The corrected total sum of squares is computed as

$$
\begin{aligned}
CSS(\text{total}) &= USS(\text{total}) - SSI \\
&= \mathbf{y}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y}}{n} \\
&= \mathbf{y}'\left[\mathbf{I}_n - \frac{\mathbf{J}_n\mathbf{J}_n'}{n}\right]\mathbf{y} \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2.
\end{aligned}
$$

The corrected model sum of squares is also computed by subtracting the $SSI$:

$$\begin{aligned} CSS(\text{model}) &= USS(\text{model}) - SSI \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y}}{n} \\ &= \mathbf{y}'\left[\mathbf{H} - \frac{\mathbf{J}_n\mathbf{J}_n'}{n}\right]\mathbf{y} \\ &= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2. \end{aligned}$$

So we see that

$$\begin{aligned} CSS(\text{total}) &= CSS(\text{model}) + SSE \\ \mathbf{y}'\left[\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} &= \mathbf{y}'\left[\mathbf{H} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

## Example: Computing Corrected Sums of Squares

The additional R code for computing the corrected sum of squares in R is given below.

```
> one = matrix(1, n) # column vec of 1s
> I_n = diag(as.vector(one)) # diagonal matrix of 1s
> css_t=t(y)%*%(I_n-(one%*%t(one))/n)%*%y
> css_m=t(y)%*%(H-(one%*%t(one))/n)%*%y
> print(css_t)

        [,1]
[1,] 15183.1

> print(css_m)

         [,1]
[1,] 5034.907

> print(css_m + sse)

        [,1]
[1,] 15183.1
```

## We may also obtain this from PROC GLM

```
proc glm data=ozone;
model personal=outdoor home time_out;
run;
```

```
                              The GLM Procedure
          Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                          Sum of
 Source                 DF        Squares    Mean Square   F Value   Pr > F

 Model                   3     5034.90667     1678.30222      9.92   <.0001

 Error                  60    10148.19129      169.13652

 Corrected Total        63    15183.09796
```

## The Intercept Only Model

Consider the following model with only an intercept:

$$\begin{aligned} \mathbf{y}_{n\times 1} &= \beta_0 + \boldsymbol{\varepsilon}_{n\times 1} \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \end{aligned}$$

It is easy to show that $\widehat{\beta}_0 = \overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, i.e.,that the estimate of $\beta_0$ is the average response. This model is sometimes called the *grand mean model*. The intercept only model predicts a constant ($\widehat{y}_i = \widehat{y}_j = \overline{y}$ for all $i, j$), and the predicted regression line has zero slope. For the ozone data, an intercept only model expects that personal exposure would be the same for all students. The estimate of the intercept is simply the mean personal exposure, so that we have $\widehat{\beta_0} = 23.55$.

Intercept-only Model

Although this model is not very useful by itself, it is very useful in model selection and testing. We will use it to answer questions like the following: "Does adding home exposure to this basic model give us any additional useful information about personal exposure?"

For the intercept only model, we have

- $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \mathbf{J}_n\widehat{\beta}_0 = \mathbf{J}_n\overline{y} = \begin{bmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{bmatrix}_{n\times 1},$

- $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \begin{bmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{bmatrix}_{n\times 1},$

- $SSE = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right)\mathbf{y},$

- $USS(\text{total}) = \mathbf{y}'\mathbf{y}$

- $USS(\text{model}) = \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'\left(\frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right)\mathbf{y},$

- $$\begin{aligned} CSS(\text{total}) &= USS(\text{total}) - SSI \\ &= \mathbf{y}' \left( \mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n' \right) \mathbf{y} = SSE, \text{ and} \end{aligned}$$

- $$\begin{aligned} CSS(\text{model}) &= USS(\text{model}) - SSI \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y} \\ &= \mathbf{y}' \left( \frac{1}{n}\mathbf{J}_n\mathbf{J}_n' \right) \mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y} \\ &= 0. \end{aligned}$$

## Example: Computing Sums of Squares for the Intercept Only Model

```
proc glm data=ozone;
model personal= /;
run;
```

```
*********************************************************************

                              The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                        Sum of
 Source                DF        Squares    Mean Square   F Value   Pr > F

 Model                  1    35481.84414    35481.84414    147.23   <.0001

 Error                 63    15183.09796      241.00155

 Uncorrected Total     64    50664.94210
```

From the SAS output, what are the values of the following?

- $SSE$:
- $USS$(total):
- $USS$(model):
- $SSI$:
- $CSS$(total):
- $CSS$(model):

What test is provided in the SAS output?

## The Null Model

An even simpler model exists. The "null" model assumes the intercept is zero and all slopes are zero so that we have no parameters in the model (except for $\sigma^2$) and thus 0 model degrees of freedom. Call

$$\boldsymbol{y} = \boldsymbol{\varepsilon}$$

the *null model*, with $p = 0$ and $\boldsymbol{\beta} = \emptyset$.

Necessarily, $\widehat{\boldsymbol{y}} = \mathbf{0}$ and $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y}$. We do not define corrected sums of squares for the null model because it does not span an intercept. For this model, we have

$$SSE \quad = \quad \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = \mathbf{y}'\mathbf{y} = USS(\text{total})$$

and $USS(\text{model}) = USS(\text{total}) - SSE = 0$.

Like the intercept only model, this model is not very interesting by itself. However, it is also useful for purposes of model selection and testing.

## Overall ANOVA Table for Multiple Regression

Typically, with all continuous predictors, $\boldsymbol{X}$ has full rank unless an error has been made. By default, for now we will assume that $\boldsymbol{X}$ less than full rank implies an error. Writing $\widehat{\boldsymbol{\beta}}$ presumes full rank $\boldsymbol{X}$, while $\tilde{\boldsymbol{\beta}}$ allows either full rank (rank$(\boldsymbol{X}) = p$) or less than full rank $\mathbf{X}$.

After fitting a model, one wishes to summarize the analysis and decide whether the variables in $\mathbf{X}$ are useful predictors (reduce the variability of $\boldsymbol{y} \mid \boldsymbol{X}$).

All regression tests considered are special cases of the GLH, which we have already studied.

Overall ANOVA Table for Model:

$\text{Wingspan}_i = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\text{Height}_i + \boldsymbol{\beta_2}\text{Waist}_i + \boldsymbol{\varepsilon_i}$

| Source | $df$ | $SS$ | $MS$ | $F_{obs}$ | $p$ |
|---|---|---|---|---|---|
| Intercept | 1 | 83.13 | | | |
| Regression (uncorrected) | 3 | 83.69 | | | |
| Regression (corrected) | 2 | 0.56 | | | |
| Error (residual) | 27 | 0.37 | | | |
| Total (uncorrected) | 30 | 84.06 | | | |
| Total (corrected) | 29 | 0.94 | | | |

*Source* indicates the variables which contribute the information.

*Degrees of freedom ($df$)* gives the dimension of the source (number of contributing variables).

Sums of squares ($SS$) are defined in the preceding section.

*Mean Square* $(MS) = SS/df$ are sums of squares adjusted for the

sample size.

$F_{obs} = MS(\text{source})/MS(\text{error})$.

Under HILE Gauss and $H_0$: $\sigma^2_{\text{source}} = \sigma^2_{\text{error}}$,

$F_{obs} \sim F(df_{\text{source}}, df_{\text{error}})$, and

$p = \Pr\{F_{obs} \geq F(df_{\text{source}}, df_{\text{error}})\}$.

Overall ANOVA Table Formulas for Full Model
(Assuming $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ Spans an Intercept)

| Source | $SS$ Form | Scalar Form | Quadratic Form |
|---|---|---|---|
| Intercept | $SSE_\emptyset - SSE_0$ | $n\overline{y}^2$ | $\boldsymbol{y}'(\frac{\mathbf{JJ}'}{n})\boldsymbol{y}$ |
| Regression (uncorrected) | $SSE_\emptyset - SSE_{p-1}$ | $\sum_{i=1}^n \widehat{y}_i^2$ | $\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}$ |
| Regression (corrected) | $SSE_0 - SSE_{p-1}$ | $\sum_{i=1}^n \widehat{y}_i^2 - n\overline{y}^2$ | $\boldsymbol{y}'(\boldsymbol{H} - \frac{\mathbf{JJ}'}{n})\boldsymbol{y}$ |
| Error (residual) | $SSE_{p-1}$ | $\sum_{i=1}^n (y_i - \widehat{y}_i)^2$ | $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ |
| Total (uncorrected) | $SSE_\emptyset$ | $\sum_{i=1}^n y_i^2$ | $\boldsymbol{y}'\boldsymbol{y}$ |
| Total (corrected) | $SSE_0$ | $\sum_{i=1}^n y_i^2 - n\overline{y}^2$ | $\boldsymbol{y}'(\boldsymbol{I} - \frac{\mathbf{JJ}'}{n})\boldsymbol{y}$ |

Corrected sums of squares and associated statistics are not defined if a model does not span an intercept. Corrected sums of squares and tests always include the intercept in models and exclude it from tests, which is generally our objective.

## Usual ("Corrected") Overall Test for Regression

Consider testing whether the predictors in $\boldsymbol{X}$ have any value. This corresponds to the hypothesis that all slopes are zero, which corresponds to no predictive contribution (explanatory value) of the covariates, except possibly the intercept.

In this test we compare the full model,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + \varepsilon_i,$$

to the reduced model,

$$y_i = \beta_0 + \varepsilon_i,$$

to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$.

$$
\begin{aligned}
F_{obs} &= \frac{MS(\text{hypothesis})}{MSE} = \frac{SSH/dfH}{SSE/dfE} \\
&= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{CSS(\text{Regression})/(p-1)}{SSE(full)/(n-p)} \,.
\end{aligned}
$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1-\alpha,\, p-1,\, n-p) = f_{crit}$.

The usual test of overall regression assumes model spans an intercept and excludes the intercept from the test.

Overall Corrected ANOVA Table Formulas for Full Model (Assuming $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ Spans an Intercept)

| Source | Scalar Form | Quadratic Form | df |
|---|---|---|---|
| Intercept | $n\overline{y}^2$ | $\boldsymbol{y}'(\frac{\mathbf{J}\mathbf{J}'}{n})\boldsymbol{y}$ | 1 |
| Regression (corrected) | $\sum_{i=1}^n \widehat{y}_i^2 - n\overline{y}^2$ | $\boldsymbol{y}'(\boldsymbol{H} - \frac{\mathbf{J}\mathbf{J}'}{n})\boldsymbol{y}$ | $p-1$ |
| Error (residual) | $\sum_{i=1}^n (y_i - \widehat{y}_i)^2$ | $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ | $n-p$ |
| Total (corrected) | $\sum_{i=1}^n y_i^2 - n\overline{y}^2$ | $\boldsymbol{y}'(\boldsymbol{I} - \frac{\mathbf{J}\mathbf{J}'}{n})\boldsymbol{y}$ | $n-1$ |

So $CSS$(total) $=$ $CSS$(model) $+$ $SSE$. To find $USS$(total) and $USS$(model), simply add $SSI$ to the corresponding corrected values (and adjust the degrees of freedom accordingly).

## Example: Conducting the "Corrected" Overall Test for Regression

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
*********************************************************************
                                The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                          Sum of
 Source                DF        Squares    Mean Square   F Value   Pr > F

 Model                  3     5034.90667     1678.30222      9.92   <.0001

 Error                 60    10148.19129      169.13652

 Corrected Total       63    15183.09796
```

## "Uncorrected" Overall Test for Regression

Consider testing whether the variables in $\boldsymbol{X}$, including the intercept, have any value as predictors.

This corresponds to the hypothesis that all slopes and the intercept are zero, which corresponds to no contribution of any predictor.

In order to do this, compare the full model,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + \varepsilon_i,$$

to the reduced model,

$$y_i = \varepsilon_i.$$

This yields $H_0$: $\beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$.

The test statistic for the uncorrected overall test is

$$F_{obs} = \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})}$$

$$= \frac{USS(\text{Regression})/p}{SSE(\text{full})/(n-p)}\,.$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1-\alpha,\, p,\, n-p) = f_{crit}$.

This test is well defined whether or not the model spans an intercept, but it rarely has scientific value.

Overall Uncorrected ANOVA Table Formulas for Full Model (Assuming $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ Spans an Intercept)

| Source | Scalar Form | Quadratic Form | df |
|---|---|---|---|
| Regression (uncorrected) | $\sum_{i=1}^{n} \widehat{y}_i^2$ | $\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}$ | $p$ |
| Error (residual) | $\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$ | $\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$ | $n - p$ |
| Total (uncorrected) | $\sum_{i=1}^{n} y_i^2$ | $\boldsymbol{y}'\boldsymbol{y}$ | $n$ |

So $USS(\text{total}) = USS(\text{model}) + SSE$, and $SSI$ is included in the regression sum of squares.

## Strength of Association

### Decomposing Response Variance

Statistical "significance" does not measure scientific importance.

We may decompose the variance into variance explained by the model and random error, and then we may ask: what fraction of $Y$ variance does $\boldsymbol{X}$ predict?

## Usual "Corrected" $R^2$

For a model spanning an intercept, define the proportion of variance in $Y$ predictable from the $X$'s, adjusted for the intercept, as

$$R_{\mathsf{c}}^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE(\text{full})}$$

$$= \frac{CSS(\text{Regression})}{CSS(\text{total})}$$

$R_c^2$ estimates $\rho_{\mathsf{c}}^2$, the population ratio of model to total variance, with $0 \leq \rho_{\mathsf{c}}^2 \leq 1$ and $0 \leq R_{\mathsf{c}}^2 \leq 1$.

*Facts about $R_c^2$:*

- $R_c^2$ is the maximum likelihood estimate of $\rho_{\mathsf{c}}^2$ under HILE Gauss.
- $R_c^2$ is biased in general: $E[R_{\mathsf{c}}^2] \geq \rho_{\mathsf{c}}^2$, with equality for $\rho_{\mathsf{c}}^2 = 1$ or $n \to \infty$.
- $R_{\mathsf{c}}^2$ is invariant with respect to full rank linear transformation of the response or predictors (including location and scale changes).
- The corrected test for overall regression,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

holds if and only if

$$H_0 : \rho_{\mathsf{c}}^2 = 0$$

is true.

## "Uncorrected" $R^2$

For any model define the proportion of variance in $Y$ predictable from the $X$'s, including the intercept if spanned, as

$$R^2_{\text{u}} = \frac{USS(\text{Regression})}{USS(\text{Regression}) + SSE(\text{full})} \ ,$$

$$= \frac{USS(\text{Regression})}{USS(\text{total})}$$

where $0 \leq R^2_{\text{u}} \leq 1$ and $0 \leq \rho^2_{\text{u}} \leq 1$.

*Facts about $R_u^2$:*

- $R_{\mathsf{u}}^2$ may vary due to any linear transformation of $Y$ or $X$'s.
- The uncorrected test for overall regression,

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0,$$

  holds if and only if

$$H_0 : \rho_{\mathsf{u}}^2 = 0$$

  is true.
- $R_u^2$ is a biased estimate of $\rho_u^2$
- Uncorrected $R^2$ is defined for any GLM, while corrected $R^2$ is defined only for models that span an intercept.
- Neither corrected nor uncorrected $R^2$ is always best, although corrected $R^2$ should be the default choice.

## Comparing Corrected and Uncorrected $R^2$

$$
\begin{aligned}
R_c^2 &= \frac{CSS(\text{Model})}{CSS(\text{Total})} \\
R_u^2 &= \frac{USS(\text{Model})}{USS(\text{Total})} = \frac{CSS(\text{Model}) + SSI}{CSS(\text{Total}) + SSI}
\end{aligned}
$$

$$0 \leq R_{\mathsf{c}}^2 \leq 1$$

$$0 \leq R_{\mathsf{u}}^2 \leq 1$$

As always, $SSI = n\overline{y}^2$.

Note: $R^2$ always increases when additional predictors are added to the model, whether or not they are practically or statistically important.

## Exercise: Computing Corrected and Uncorrected $R^2$

Compute corrected and uncorrected $R^2$ for the ozone data.

## Adjusted $R^2$

One potential problem with using $R^2$ (either corrected or uncorrected) to assess model adequacy is that $R^2$ never decreases when you add additional predictors to the model. (Even if a predictor is meaningless, $R^2$ will never be smaller when it is added.) Although we will later learn how to conduct hypothesis tests about whether an increase in $R^2$ is statistically significant, some investigators prefer to use the *adjusted* $R^2$, defined as

$$R^2{}_{adj} = 1 - \frac{SSE/(n-r)}{CSS(\text{Total})/(n-1)},$$

which is adjusted for the degrees of freedom. It will only increase on adding a variable to the model if the variable reduces the mean square for error. The fact that adjusted $R^2$ penalizes us for adding terms that are not useful makes it helpful in evaluating and comparing a set of candidate regression models.

## Next: Testing

*Reading Assignment:*

- Muller and Fetterman, Chapter 5: "Testing Hypotheses in Multiple Regression" (Required)

# Lecture 8: Testing Hypotheses in Multiple Regression

*Reading Assignment:*

- Muller and Fetterman, Chapter 5: "Testing Hypotheses in Multiple Regression"

After fitting a model, one seeks to draw inferences about parameters. Correlations and confidence intervals measure scientific importance, while tests and p-values assess statistical "significance". The two concepts are not necessarily the same!

## Review of GLH Concepts

- Assume HILE Gauss and let $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- We estimate the $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ or $\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$.
- Compute $\widehat{\boldsymbol{y}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}}$, $SSE = (\boldsymbol{y} - \widehat{\boldsymbol{y}})'(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \mathbf{y}'\left[\mathbf{I} - \mathbf{H}\right]\mathbf{y}$, and $\widehat{\sigma}^2 = MSE = \frac{SSE}{dfE}$, with $r = \text{rank}(\boldsymbol{X})$ and $dfE = n - r$.
- Define $\boldsymbol{C}_{a\times p}$, which implies $\boldsymbol{\theta} = \boldsymbol{C}\boldsymbol{\beta}$, and state the GLH as $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. If $\boldsymbol{X}$ is less than full rank, check estimability.
- Estimate secondary parameters: $\widehat{\boldsymbol{\theta}} = \boldsymbol{C}\tilde{\boldsymbol{\beta}}$.
- Compute $\boldsymbol{M} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'$ and

$$SSH = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

- Compute $F_{obs} = \frac{SSH/dfH}{SSE/dfE}$ for $dfH = a$. reject $H_0$ if $F_{obs} > F_F^{-1}(1 - \alpha, a, n - r)$.

Although all the tests we consider fall into the framework of the General Linear Hypothesis (GLH), we will discuss several different subclasses of hypothesis tests in the GLH framework, including the following.

1. *Overall tests*, which measure the contribution of the entire set of predictors
2. *Addition of one variable tests*, which measure the contribution of one variable beyond others
3. *Intercept tests*, which measure the value of the intercept in predicting the response
4. *Addition of a group of variables tests*, which measure the contribution of a group of variables beyond that of others
5. GLH tests, which include the above tests and others not included in the above categories.

*Calculating Test Statistic:*

A GLM test statistic can be expressed in terms of $SSE$'s (and $SSR$'s) or $\{(\boldsymbol{X}'\boldsymbol{X})^{-1}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2, \boldsymbol{C}, \boldsymbol{\theta}_0\}$. Under HILE Gauss, the likelihood ratio test statistic follows an $F$ distribution.

All tests compare two models: the full model and the reduced model (this is the basic idea of likelihood ratio tests, called the *likelihood ratio principle*).

We say that the reduced model is *nested* in the full model (so that its parameters are a subset of the parameters in the full model). Sometimes, investigators wish to compare non-nested models. (Example: An investigator wants to compare a model that uses weight as a continuous predictor to a model that uses indicator variables for overweight, underweight, or normal weight.) This is much more difficult and involves model-selection criteria other than F-tests. We

will only consider *nested* models for the present.

It is important to recognize that tests decompose a fixed amount of variance, $\sigma_y^2$, so that sums of squares are not created or erased but are simply moved into different locations (from $SSE$ to $SSR$ or vice versa). An important consequence of this fact is that $SSR$ increases $\Leftrightarrow$ $SSE$ decreases.

## Choosing an Error Term

In many cases, a series of nested models are compared. Tests are conducted for each model in the series, comparing that model to another.

An important question is "What error term should be used?".

- It is safest to use the error term from the largest model in the entire series instead of the larger of any given pair because $SSE/(n-r)$ from the largest model "guarantees" (by assumption) an unbiased estimate of $\sigma^2$.
- If terms beyond those in the smaller model of a pair are unimportant, then the $SSE$ from the smaller model allows a more powerful test because the error df are $(n-r+d)$ rather than $(n-r)$, where $d$ indicates the number predictors by which the models differ.

- If terms beyond those in the smaller model of a pair are important, but we use the $SSE$ from the smaller, inadequate model, then we will inflate (bias upwards) the estimate of $\sigma^2$. This may substantially reduce power.
- A model that is too small (under-fitting) gives biased $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ and possibly leads to a large power loss for any $n$.
- A model that is too large gives unbiased $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ and usually only a small power loss that goes to zero as $n$ goes to $\infty$.
- Because $(n - r + d)/(n - r)$ should be near 1.0 with a sufficient amount of data, the choice of error term should not really matter.
- As a general rule, we recommend using $\widehat{\sigma}^2$ from the largest model in a series. This largest model is often called (somewhat loosely) the *full model*.

## Comparing Two Models

If two models differ only by the addition or deletion of one or more variables, then

$$\begin{aligned} USS(\text{larger}) &\geq USS(\text{smaller}) \\ CSS(\text{larger}) &\geq CSS(\text{smaller}) \\ SSE(\text{larger}) &\leq SSE(\text{smaller}), \end{aligned}$$

because the larger model explains more of the variability in the data and thus has a larger SS model (and smaller SSE) than the smaller model. The SS Total of the two models should be identical.

Note that $CSS$ may not be defined for all models.

## Test Class I: Overall Tests

### Corrected Overall Test

The *corrected overall test* compares the full model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{ip-1} + \varepsilon_i,$$

to the intercept-only model,

$$y_i = \beta_0 + \varepsilon_i.$$

If the model does not span an intercept this test is not defined.

This test corresponds to

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

and

$$H_0 : \rho_{\mathsf{c}}^2 = 0.$$

Let $SSE(\text{reduced}) = SSE(\beta_0)$ be the sum of squares for error for the intercept-only model, and let $SSE(\text{full}) = SSE(\beta_0, \beta_1, \ldots, \beta_{p-1})$ be the $SSE$ for the full model. (Generally, $SSE(\beta_i, \beta_j)$ is the $SSE$ for a model including the two parameters $\beta_i$ and $\beta_j$.) Then the sum of squares for the hypothesis is given by

$$\begin{aligned} SSH &= SSE(\beta_0) - SSE(\beta_0, \beta_1, \ldots, \beta_{p-1}) \\ &= SSE(\text{reduced model}) - SSE(\text{full model}) \\ &= CSS(\text{full model}) - CSS(\text{reduced model}) \\ &= CSS(\text{full model}) - 0. \end{aligned}$$

The $F$ statistic is

$$
\begin{aligned}
F_{obs} &= \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\dfrac{SSE(\beta_0) - SSE(\beta_0, \beta_1, \ldots \beta_{p-1})}{dfE(\beta_0) - dfE(\beta_0, \beta_1, \ldots \beta_{p-1})}}{SSE(\beta_0, \beta_1, \ldots \beta_{p-1})/dfE(\beta_0, \beta_1, \ldots \beta_{p-1})} \\
&= \frac{CSS(\beta_0, \ldots, \beta_{p-1})/(p-1)}{SSE(\beta_0, \ldots, \beta_{p-1})/(n-p)},
\end{aligned}
$$

for full rank $\boldsymbol{X}$.

The hypothesis is rejected if

$$F_{obs} > F_F^{-1}(1-\alpha, p-1, n-p).$$

We may also obtain the same test via the GLH formulation with

$$F_{obs} \quad = \quad \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/dfE},$$

where

- $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'M^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is the $SSH$,
- $dfH = a$ is the hypothesis degrees of freedom or the number of rows of $\boldsymbol{\theta}$ (unless we pick a redundant $\boldsymbol{\theta}$),
- $\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$ is the $SSE$, and
- $dfE = \text{rank}(\mathbf{I} - \mathbf{H})$ is the degrees of freedom for error.

To conduct the corrected overall test, we pick

$$\boldsymbol{C} = \begin{bmatrix} \mathbf{0}_{p-1} & \mid \boldsymbol{I}_{p-1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{bmatrix}_{(p-1)\times p}$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$, which imply that $\boldsymbol{\theta} = [\beta_1, \ldots \beta_{p-1}]'$ and $a = p - 1$.

The following R and SAS code may be used to construct the corrected overall test for the ozone data, which is a test of $H_0 : \beta_{OUTDOOR} = \beta_{HOME} = \beta_{TIMEOUT} = 0$.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)
```

```
  (Intercept)  outdoor  home time_out
1           1 35.87771 22.29     0.57
2           1 43.79189 13.97     0.90
3           1 49.81255 18.96     0.55
4           1 34.37366 22.27     0.17
5           1 45.95496 23.40     0.00
6           1 64.76558 39.62     0.30
```

```
> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
```

```
> I_n = diag(rep(1, n))
> H=X%*%solve(t(X) %*% X) %*% t(X)
> sse = t(y) %*%(I_n - H)%*% y # SSE
> mse=sse/df; # MSE
> C=matrix(c(0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1), nrow = 3, byrow = T)
> print(C)
     [,1] [,2] [,3] [,4]
[1,]    0    1    0    0
[2,]    0    0    1    0
[3,]    0    0    0    1
> M=C%*%solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(thetahat)%*%solve(M)%*%thetahat
> msh=ssh/nrow(thetahat)
> f_obs=msh/mse
> p=1-pf(f_obs,3,60)
> print(f_obs)
         [,1]
[1,] 9.922767
> print(p)
```

```
              [,1]
[1,] 2.094332e-05
```

We may also obtain the same values in SAS after running the following

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
```

```
                          The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                        Sum of
Source                      DF         Squares    Mean Square   F Value

Model                        3      5034.90667     1678.30222      9.92

Error                       60     10148.19129      169.13652

Corrected Total             63     15183.09796

                     Source                  Pr > F

                     Model                   <.0001

                     Error

                     Corrected Total
```

We reject the null hypothesis that outdoor concentration of ozone, home concentration of ozone, and proportion of time spent outdoors have no effect on personal ozone exposure. We conclude that at least one of our predictors (outdoor concentration of ozone, home concentration of ozone, or proportion of time spent outdoors) is related to personal ozone exposure.

## Heuristics

For data that are exactly normal, this $F$ test is a *likelihood ratio test*. You will learn more about the likelihood ratio test in BIOS 661, so our discussion will be brief.

Our general approach is to fit two models: one full model containing all the parameters of interest, and one smaller model with only an intercept. Then, we compare their maximized likelihoods (or log-likelihoods).

For a fixed value of $\sigma^2$, the maximized log-likelihood of the intercept-only model is

$$\max \log L(\beta_0) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{SSE(\beta_0)}{\sigma^2},$$

and the maximized log-likelihood of the larger model is

$$\max \log L(\beta_0, \ldots, \beta_{p-1}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{SSE(\beta_0, \ldots, \beta_{p-1})}{\sigma^2}.$$

To compare the log-likelihoods, we calculate minus twice their difference (denoted $-2\log\lambda$), which is given by

$$\frac{SSE(\beta_0) - SSE(\beta_0, \ldots, \beta_{p-1})}{\sigma^2}.$$

We see that this criterion is used to judge whether the additional predictors associated with $(\beta_1, \ldots, \beta_{p-1})$ result in a significant reduction in the SSE. This likelihood ratio criterion, $-2\log\lambda$, has an exact $F$ distribution when our data are exactly normal. In BIOS 661, you will learn more about the asymptotic distribution of this criterion, useful when the data are *not* exactly normal.

## Uncorrected Overall Test (rarely used)

The *uncorrected overall test* is a test of whether all parameters (including an intercept, if there is one) are equal to zero. This test is defined whether or not the model includes an intercept. It compares the null model,

$$y_i = \varepsilon_i,$$

with the full model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

and tests both

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

and

$$H_0 : \rho_{\mathsf{u}}^2 = 0.$$

The test statistic and p-value are location dependent.

The $F$ statistic is

$$
\begin{aligned}
F_{obs} &= \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\dfrac{SSE(\text{null}) - SSE(\beta_0, \beta_1, \ldots \beta_{p-1})}{dfE(\text{null}) - dfE(\beta_0, \beta_1, \ldots \beta_{p-1})}}{SSE(\beta_0, \beta_1, \ldots \beta_{p-1})/dfE(\beta_0, \beta_1, \ldots \beta_{p-1})} \\
&= \frac{USS(\beta_0, \ldots, \beta_{p-1})/p}{SSE(\beta_0, \ldots, \beta_{p-1})/(n-p)},
\end{aligned}
$$

for full rank $\boldsymbol{X}$, the hypothesis is rejected if $F_{obs} > F_F^{-1}(1-\alpha, p, n-p)$.

We may also obtain the same test via the GLH formulation with

$$F_{obs} \quad = \quad \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/dfE},$$

where $dfH = p$ and $dfE = n - p$ when $\mathbf{X}$ is full rank.

To conduct the uncorrected overall test, we pick

$$\boldsymbol{C} = \begin{bmatrix} \mathbf{I}_p \end{bmatrix} = \begin{bmatrix} 1 & & 0 \\ \vdots & \ddots & \\ 0 & & 1 \end{bmatrix}_{p \times p}$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$. Thus $\boldsymbol{\theta} = [\beta_0, \beta_1, \ldots \beta_{p-1}]'$ and $a = p$.

## Model Pools

### Added-Last Pool

With $p$ parameters, $2^p$ distinct models can be defined. We usually only consider subsets formed by some rule. For added-last tests, model $j$ has $\beta_j$ deleted and contains $p-1$ parameters.

An *added-last test* compares the full model to a reduced model that is obtained by deleting the $j$th variable from the full model. In this subset, there exist $p$ reduced models (corresponding to deleting each of the $p$ parameters). Each reduced model has $p-1$ parameters, and the full model has $p$ parameters. The model pool for added-last testing is provided below.

Added-Last Model Pool

| Model | | | | | |
|---|---|---|---|---|---|
| $-0$ | $y_i =$ | $\beta_1 x_{i1}$ | $+\beta_2 x_{i2}$ | $\cdots$ | $\cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$ |
| $-1$ | $y_i = \beta_0$ | | $+\beta_2 x_{i2}$ | $\cdots$ | $\cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$ |
| $\vdots$ | | | | | |
| $-j$ | $y_i = \beta_0 + \beta_1 x_{i1}$ | $\cdots$ | $+\beta_{j-1} x_{ij-1}$ | | $\cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$ |
| $\vdots$ | | | | | |
| full | $y_i = \beta_0 + \beta_1 x_{i1}$ | $\cdots$ | $+\beta_{j-1} x_{ij-1} + \beta_j x_{ij}$ | $\cdots$ | $+ \beta_{p-1} x_{ip-1} + \varepsilon_i$ |

**Added-In-Order Pool**

An *added-in-order test* compares parameters added in sequence. In this subset, there exist $p+1$ models (the null model and $p$ models corresponding to adding each of the $p$ parameters). The $(p+1)$th model is the full model and has $p$ parameters, the $p$th model has one fewer parameter, the $(p-1)$th model has two fewer parameters, etc.

The model pool for added-in-order testing is provided below.

Added-in-Order Model Pool

| Model | | |
|---|---|---|
| $\emptyset$ | $y_i =$ | $\varepsilon_i$ |
| $0$ | $y_i = \beta_0$ | $+\varepsilon_i$ |
| $1$ | $y_i = \beta_0 + \beta_1 x_{i1}$ | $+\varepsilon_i$ |
| $\vdots$ | | |
| $j-1$ | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{j-1} x_{ij-1}$ | $+\varepsilon_i$ |
| $j$ | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij}$ | $+\varepsilon_i$ |
| $\vdots$ | | |
| $p-1$ | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{j-1} x_{ij-1} + \beta_j x_{ij} \ldots + \beta_{q-1} x_{iq-1}$ | $+\varepsilon_i$ |

Model $p-1$ is the full model.

## Test Class 2: Addition of One Variable

### Added-Last Test

The *added-last test* seeks to assess the usefulness of one predictor, above and beyond all others. An added-last test for variable $j$ compares the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{j-1} x_{ij-1} + \beta_{j+1} x_{ij+1} + \ldots + \beta_{p-1} x_{ip-1} + \varepsilon_i$$

to the full model and corresponds to testing the hypothesis $H_0 : \beta_j = 0$ when $\beta_j$ is the final variable added to the model.

In this test class, there are $(p-1)$ reduced models (eliminating each of the $(p-1)$ non-intercept parameters) with $(p-1)$ parameters each (including the intercept). We compare each reduced model to the full model to produce an added-last-test for each predictor.

Added-last tests do not depend on the order of variables in the model. Added-last tests are printed in SAS as the Type III SS tests.

Denote the sum of squares for error for the reduced model (without the $j$th variable) by

$$SSE(\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_{p-1}).$$

Then the F statistic is

$$\begin{aligned}
F_{obs} &= \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\frac{SSE(\beta_0,\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_{p-1}) - SSE(\beta_0,\ldots,\beta_{p-1})}{dfE(\beta_0,\beta_1,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_{p-1}) - dfE(\beta_0,\ldots,\beta_{p-1})}}{SSE(\beta_0, \ldots, \beta_{p-1})\big/dfE(\beta_0, \ldots, \beta_{p-1})} \\
&= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}\big/dfE},
\end{aligned}$$

where

$$\mathbf{C} = \begin{bmatrix} 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}_{1 \times p}$$

has a 1 in the position corresponding to the variable we wish to test, and 0 elsewhere. If we let $\theta_0 = 0$, then we have $\theta = \beta_j$ and $a = 1$.

$T_{obs} = F_{obs}^{\frac{1}{2}}$ follows a Student's $T$ distribution with $(n - p)$ df if $\boldsymbol{X}$ is full rank. Many regression programs report $T_{obs}$ as a test of the $j$th regression coefficient equaling zero, so the usual $T$ test for a regression coefficient provides an added last test.

We reject the hypothesis if $F_{obs} \geq F_F^{-1}(1 - \alpha, 1, n - p)$.

SAS code and output for an added-last test of the variable $O_{OUTDOOR}$ is provided.

```
> C=matrix(c(0, 1, 0, 0), nrow = 1, byrow = T)
> print(C)

     [,1] [,2] [,3] [,4]
[1,]    0    1    0    0

> M=C%*%solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(thetahat)*solve(M)%*%thetahat
> msh=ssh/nrow(thetahat)
> f_obs=msh/mse
> p=1-pf(f_obs,1,60)
> print(f_obs)

         [,1]
[1,] 1.022314

> print(p)

          [,1]
[1,] 0.3160309
```

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
```

```
                          The GLM Procedure

Source                      DF      Type I SS    Mean Square   F Value

outdoor                      1    2419.043105    2419.043105     14.30
home                         1    2086.106656    2086.106656     12.33
time_out                     1     529.756909     529.756909      3.13

                    Source                 Pr > F

                    outdoor                0.0004
                    home                   0.0009
                    time_out               0.0818


Source                      DF    Type III SS    Mean Square   F Value

outdoor                      1     172.910655     172.910655      1.02
home                         1    2208.421320    2208.421320     13.06
time_out                     1     529.756909     529.756909      3.13

                    Source                 Pr > F

                    outdoor                0.3160
                    home                   0.0006
```

```
                    time_out                0.0818

                                     Standard
  Parameter          Estimate            Error    t Value    Pr > |t|

  Intercept        3.78348593       4.34205547       0.87      0.3870
  outdoor          0.09142005       0.09041683       1.01      0.3160
  home             0.59543659       0.16478332       3.61      0.0006
  time_out        13.64453832       7.70973105       1.77      0.0818
```

Again, note that the $t$ value squared is equal to the $F$ statistic for the added-last test (Type III SS). We fail to reject the null hypothesis and conclude that outdoor ozone concentration does not make a significant contribution to personal exposure above and beyond the intercept, home ozone, and the proportion of time spent outdoors.

## Caution: Added-Last Tests

Suppose you want to estimate a person's weight given the length of their legs. (Legs account for 30-35% of human body weight.) First, you fit the model

$$E(\text{weight}) = \beta_0 + \beta_1 \text{ right leg length } .$$

The results of this model fit are provided below.

```
proc glm;
model weight=rleg;
run;
*************************************************************************
                          The GLM Procedure

Dependent Variable: weight

                                          Sum of
Source                      DF          Squares     Mean Square    F Value
Model                        1      3627.670184     3627.670184     125.75
Error                       98      2827.099916       28.847958
Corrected Total             99      6454.770100

                    Source                    Pr > F
```

```
                    Model                    <.0001


         R-Square     Coeff Var      Root MSE    weight Mean
         0.562016      6.969661      5.371030       77.06300

Source                      DF      Type I SS    Mean Square   F Value
rleg                         1    3627.670184    3627.670184    125.75

                    Source                   Pr > F
                    rleg                     <.0001

Source                      DF    Type III SS    Mean Square   F Value
rleg                         1    3627.670184    3627.670184    125.75

                    Source                   Pr > F
                    rleg                     <.0001

                                     Standard
  Parameter         Estimate            Error    t Value    Pr > |t|
  Intercept     -3.731074382       7.22481246      -0.52      0.6067
  rleg           1.008948511       0.08997309      11.21      <.0001
```

The added-last test, given by the "Type III" results, leads us to conclude that right leg length is significantly related to weight.

Now consider the results from fitting the model

$$E(\text{weight}) = \beta_0 + \beta_1 \text{ right leg length } + \beta_2 \text{ left leg length.}$$

```
proc glm;
model weight=rleg lleg;
run;
*********************************************************************
                          The GLM Procedure

Dependent Variable: weight

                                          Sum of
Source                      DF          Squares     Mean Square    F Value
Model                        2      3698.009886     1849.004943      65.06
Error                       97      2756.760214       28.420208
Corrected Total             99      6454.770100

                    Source                  Pr > F
                    Model                   <.0001

         R-Square     Coeff Var      Root MSE    weight Mean
         0.572911      6.917795      5.331061       77.06300

Source                      DF       Type I SS     Mean Square    F Value
```

```
rleg                        1    3627.670184    3627.670184    127.64
lleg                        1      70.339702      70.339702      2.47

                    Source                 Pr > F
                    rleg                   <.0001
                    lleg                   0.1189

Source                     DF    Type III SS    Mean Square   F Value
rleg                        1    88.17311509    88.17311509      3.10
lleg                        1    70.33970223    70.33970223      2.47

                    Source                 Pr > F
                    rleg                   0.0813
                    lleg                   0.1189

                                     Standard
  Parameter         Estimate            Error    t Value    Pr > |t|
  Intercept     -1.828644701       7.27229423      -0.25      0.8020
  rleg           9.433109580       5.35550484       1.76      0.0813
  lleg          -8.447148369       5.36937194      -1.57      0.1189
```

The added-last tests of right leg length (p=0.08) and left leg length (0.12) indicate that neither significantly predicts weight! This occurs because after you adjust for the effect of the length of one leg, the other provides no additional useful information. Such situations arise often in biostatistics, especially when predictors may be highly correlated.

With an added-in-order test ("Type I" in SAS), which we will discuss next, instead of evaluating each variable in the presence of all others, we will evaluate variables in a fixed order. Using that type of test for the leg data, you find that the leg added first (here, the right leg) does significantly predict weight, while the addition of the second leg to the model does not sharpen our prediction over a model already containing one leg. Due to the high correlation of left and right leg lengths (here, greater than 0.99), this result is not surprising.

## Added-in-Order Test

The *added-in-order test* seeks to assess the contribution of predictor $j$ above and beyond all of the preceding $j-1$ predictors (without the $j+1$, $j+2$, etc.predictors in the model).

Added-in-order SS are mutually exclusive and together exhaustive pieces of the model SS (i.e.,if you add the SS for each predictor, the resulting sum is the model SS). Results do depend on order of inclusion except when all predictors are uncorrelated. If all predictors are uncorrelated, then added-last SS coincide with added-in-order.

Added-in-order tests are available from SAS type I SS. The sum of squares for the hypothesis for the added-in-order test for the variable $j$ is given by

$$\begin{aligned} SSH &= SSE(\beta_0, \ldots, \beta_{j-1}) - SSE(\beta_0, \ldots, \beta_{j-1}, \beta_j) \\ &= SSE(\text{smaller}) - SSE(\text{larger}). \end{aligned}$$

The recommended $F$ statistic is given by

$$
\begin{aligned}
F_{obs} &= \frac{\dfrac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\frac{SSE(\beta_0,\beta_1,\ldots,\beta_{j-1}) - SSE(\beta_0,\ldots,\beta_j)}{dfE(\beta_0,\beta_1,\ldots,\beta_{j-1}) - dfE(\beta_0,\ldots,\beta_j)}}{SSE(\beta_0,\ldots,\beta_{p-1})\big/(n-p)} \\
&= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' M^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}\big/dfE},
\end{aligned}
$$

where

$$
\mathbf{C} = \begin{bmatrix} 0 & \ldots & 0 & 1 \end{bmatrix}_{1\times(j+1)}
$$

and multiplies the $\boldsymbol{\beta}_{(j+1)\times 1}$ vector corresponding to the larger model. If we let $\theta_0 = 0$, then we have $\theta = \beta_j$ and $a = 1$.

(Although it is possible to use the $MSE$ from the larger of the two models being compared as the denominator of the $F$ test, the $MSE$ from the full model is preferred.)

## Test Class 3: Tests of the Intercept

### General Features

While tests of the intercept are generally not recommended, they may be obtained by treating the intercept variable (a column of 1's) just like any other variable in Class 2 (Addition of One Variable Tests).

Muller and Fetterman discuss examples of situations in which tests of the intercept are valid.

## Test Class 4: Addition of a Group of Variables

The tests in this class are conducted just like tests in Class 2, with the exception that now we are testing more than one variable (and thus have more than one degree of freedom for the hypothesis).

### Group Added-Last Tests

We assume predictors fall into two groups $\{X_1, \ldots X_{g_1}\}$ and $\{X_{g_{1+1}}, \ldots X_{p-1}\}$, with $g_1$ variables in the first group and $g_2 = [(p-1) - g_1]$ variables in the second group. Note that $p = 1 + g_1 + g_2$.

Group Added-Last Pool

| Model | |
|---|---|
| 1 | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{g_1} x_{ig_1} \qquad + \varepsilon_i$ |
| 2 | $y_i = \beta_0 \qquad + \beta_{g_1+1} x_{ig_1+1} \ldots + \beta_{p-1} x_{ip-1} + \varepsilon_i$ |
| 3 | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{g_1} x_{ig_1} + \beta_{g_1+1} x_{ig_1+1} \ldots + \beta_{p-1} x_{ip-1} + \varepsilon_i$ |

The added-last test of the first group compares model 2 and model 3, with corresponding hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{g_1} = 0.$$

The added last test of the second group compares model 1 and model 3, with corresponding hypothesis

$$H_0 : \beta_{g_1+1} = \beta_{g_1+2} = \ldots = \beta_{p-1} = 0.$$

Always compare the appropriate reduced model to the full model. The methods generalize to any number of groups.

These tests are often called *lack-of-fit* tests and measure the adequacy of a smaller model compared to a larger model.

The F statistic for an added-last test of the first group is given by

$$
\begin{aligned}
F_{obs} &= \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\dfrac{SSE(\beta_{g_1+1}, \beta_{g_1+2}, \ldots, \beta_{p-1}) - SSE(\beta_0, \ldots, \beta_{p-1})}{dfE(\beta_{g_1+1}, \beta_{g_1+2}, \ldots, \beta_{p-1}) - dfE(\beta_0, \ldots, \beta_{p-1})}}{SSE(\beta_0, \ldots, \beta_{p-1})/(n-p)} \\
&= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/dfE},
\end{aligned}
$$

where

$$
\mathbf{C} = \begin{bmatrix} \mathbf{0_{g_1 \times 1}} & \mathbf{I}_{g_1} & \mathbf{0_{g_1 \times g_2}} \end{bmatrix}
$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$, which implies $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_{g_1})$ and $a = g_1$.

## Exercise: Computing a Group Added Last Test

Given the following SAS code, compute the group added-last test of no effect of outdoor ozone concentration and home ozone concentration for the ozone data.

```
proc glm data=ozone;
model personal=outdoor home time_out;
run;
proc glm data=ozone;
model personal=time_out;
run;
```

```
*************************************************************************

                          The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                          Sum of
Source                      DF          Squares     Mean Square    F Value
Model                        3       5034.90667      1678.30222       9.92
Error                       60      10148.19129       169.13652
Corrected Total             63      15183.09796
```

```
                    Source                  Pr > F
                    Model                   <.0001

        R-Square     Coeff Var      Root MSE    personal Mean
        0.331613      55.23389      13.00525         23.54578

                         The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                          Sum of
Source                      DF          Squares    Mean Square   F Value
Model                        1        574.49202      574.49202      2.44
Error                       62      14608.60594      235.62268
Corrected Total             63      15183.09796

                    Source                  Pr > F
                    Model                   0.1235


        R-Square     Coeff Var      Root MSE    personal Mean
        0.037838      65.19217      15.35001         23.54578
```

## Group Added-In-Order Tests

The *group added-in-order test* seeks to assess the usefulness of a group of predictors above and beyond that of preceding predictors.

Group Added-in-Order Pool

| Model | | |
|---|---|---|
| 0 | $y_i = \beta_0$ | $+\varepsilon_i$ |
| 1 | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{g_1} x_{ig_1}$ | $+\varepsilon_i$ |
| 2 | $y_i = \beta_0 + \beta_1 x_{i1} \ldots + \beta_{g_1} x_{ig_1} + \beta_{g_1+1} x_{ig_1+1} \ldots + \beta_{p-1} x_{ip-1}$ | $+\varepsilon_i$ |

An added-in-order test of the first group compares models 0 and 1, with corresponding hypothesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{g_1} = 0.$$

Listing order strongly affects the tests except when all predictors are uncorrelated (then added-last tests coincide with added-in-order tests).

Added-in-order sums of squares are mutually exclusive and together exhaustive, while added-last sums of squares typically does not overlap.

The F statistic for an added-in-order test of the first group is given by

$$
\begin{aligned}
F_{obs} &= \frac{\dfrac{SSE(\text{smaller}) - SSE(\text{larger})}{dfE(\text{smaller}) - dfE(\text{larger})}}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{\dfrac{SSE(\beta_0) - SSE(\beta_0, \ldots, \beta_{g_1})}{dfE(\beta_0) - dfE(\beta_0, \ldots, \beta_{g_1})}}{SSE(\beta_0, \ldots, \beta_{p-1})/(n-p)} \\
&= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/dfE},
\end{aligned}
$$

where

$$
\mathbf{C} = \begin{bmatrix} \mathbf{0}_{\mathbf{g_1} \times \mathbf{1}} & \mathbf{I}_{g_1} \end{bmatrix}
$$

and $\boldsymbol{\theta}_0 = \mathbf{0}$, which implies $\boldsymbol{\theta} = (\beta_1, \ldots, \beta_{g_1})$ and $a = g_1$.

As with a single variable test, the in-order test for the last group always coincides with the corresponding added-last test.

## Exercise: Computing a Group Added-In-Order Test

Given the additional SAS code and output below, compute the group added-in-order test of no effect of outdoor ozone concentration and home ozone concentration for the ozone data.

```
proc glm data=ozone;
model personal=/;
run;
proc glm data=ozone;
model personal=outdoor home;
run;
```

```
*************************************************************************

                          The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                          Sum of
Source                      DF          Squares     Mean Square    F Value
Model                        1      35481.84414     35481.84414     147.23
Error                       63      15183.09796       241.00155
Uncorrected Total           64      50664.94210
```

```
                    Source                  Pr > F
                    Model                   <.0001
                    Error
                   Uncorrected Total

                The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                          Sum of
Source                      DF          Squares    Mean Square   F Value
Model                        2       4505.14976     2252.57488     12.87
Error                       61      10677.94820      175.04833
Corrected Total             63      15183.09796

                    Source                  Pr > F
                    Model                   <.0001

        R-Square     Coeff Var      Root MSE    personal Mean
        0.296721      56.19089      13.23058         23.54578

Source                      DF       Type I SS    Mean Square   F Value
outdoor                      1      2419.043105    2419.043105     13.82
home                         1      2086.106656    2086.106656     11.92

                    Source                  Pr > F
```

```
                    outdoor                 0.0004
                    home                    0.0010

Source                     DF    Type III SS    Mean Square   F Value
outdoor                     1     240.247837     240.247837      1.37
home                        1    2086.106656    2086.106656     11.92

                    Source                  Pr > F
                    outdoor                 0.2459
                    home                    0.0010
```

## Test Class 5: Other GLH Tests

As we have seen so far, the GLH test is a powerful tool to test many hypotheses.

**Example: Laboratory Instrument Validation**

A hospital laboratory has obtained a new machine to measure CO bound to hemoglobin. Investigators wish to know if the new machine is producing the same readings as the old one.

If so, the predicted regression line should be a line with slope one through the origin.

Ideally,

$$
\begin{aligned}
\mathsf{new}_i &= \mathsf{old}_i + \varepsilon_i \\
&= 0 + 1 \cdot \mathsf{old}_i + \varepsilon_i \\
&= \beta_0 + \beta_1 \mathsf{old}_i + \varepsilon_i \\
\textbf{new} &= \begin{bmatrix} \mathbf{1} & \textbf{old} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\varepsilon}
\end{aligned}
$$

The equivalence of the new and old machines implies $\beta_0 = 0$ and $\beta_1 = 1$. So our null hypothesis is

$$H_0\colon \quad \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Setting $\boldsymbol{C} = \boldsymbol{I}_2$ yields $\boldsymbol{\theta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$.

Choosing $\boldsymbol{\theta}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ defines a GLH which compares two models,

$$\begin{array}{lll} 1 & \mathsf{new}_i = & \mathsf{old}_i + \varepsilon_i \\ 2 & \mathsf{new}_i = \beta_0 + \beta_1 \cdot \mathsf{old}_i & + \varepsilon_i \,, \end{array}$$

with test statistic

$$F_{obs} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/dfH}{\mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/dfE},$$

where $dfH = a = 2$.

We note that PROC GLM will only test a GLH with $\boldsymbol{\theta}_0 = \mathbf{0}$. In order to conduct this test, we must either use R or PROC REG with the TEST statement.

## Multiple Testing

In multiple regression, we may wish to conduct $p-1$ tests (one for each predictor in the model). This multiple testing often inflates $\alpha$ above the desired nominal level, which is generally $\alpha = 0.05$.

The Bonferroni inequality says that

$$\begin{aligned} \Pr\{\mathsf{A}_1 \cup \mathsf{A}_2\} &= \Pr\{\mathsf{A}_1\} + \Pr\{\mathsf{A}_2\} - \Pr\{\mathsf{A}_1 \cap \mathsf{A}_2\} \\ \Pr\{\mathsf{A}_1 \cup \mathsf{A}_2\} &\leq \Pr\{\mathsf{A}_1\} + \Pr\{\mathsf{A}_2\} \ . \end{aligned}$$

If $\Pr\{\mathsf{A}_1\} = \Pr\{\mathsf{A}_2\} = \alpha_k$, then $\Pr\{\mathsf{A}_1 \cup \mathsf{A}_2\} \leq 2\alpha_k$.

More generally

$$\Pr\{\mathsf{A}_1 \cup \mathsf{A}_2 \cup \mathsf{A}_3 \ \ldots \ \cup \mathsf{A}_K\} \leq K\alpha_k \ .$$

To ensure an upper bound of $\alpha$ overall, we use $\alpha_k = \alpha/K$.

For independent events (this rarely applies to multiple comparisons!)

$$\Pr\{\mathsf{A}_1 \cup \mathsf{A}_2 \cup \mathsf{A}_3 \ \ldots \ \cup \mathsf{A}_K\} = 1 - (1-\alpha_k)^K \ .$$

The Bonferroni correction is accurate for small $K$, is almost always accurate for independent events, but is least accurate for highly correlated events. The poor accuracy results in less powerful tests, but the Bonferroni correction does guarantee control of the type I error rate.

## Exercise: Bonferroni Correction

Suppose we will test 10 predictors and want an overall $\alpha = 0.05$. At which level should we conduct each test?

## Interaction

$X_1$ and $X_2$ interact (in predicting $Y$) if one must know the level of $X_1$ in order to describe relationship of $Y$ to $X_2$ (and hence must know the level of $X_2$ in order to describe relationship of $Y$ to $X_1$).

For example, one cancer treatment may be beneficial to patients with a certain genotype of a gene but not to other patients.

## Interaction Model Pool

| Model | | |
|---|---|---|
| 1 | $y_i = \beta_0 \qquad\qquad\qquad\qquad + \beta_3 x_{i1} x_{i2} + \varepsilon_i$ | *avoid* |
| 2 | $y_i = \beta_0 \qquad\qquad + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$ | *avoid* |
| 3 | $y_i = \beta_0 + \beta_1 x_{i1} \qquad\qquad + \beta_3 x_{i1} x_{i2} + \varepsilon_i$ | *avoid* |
| 4 | $y_i = \beta_0 \qquad\qquad\qquad\qquad\qquad\qquad + \varepsilon_i$ | |
| 5 | $y_i = \beta_0 + \beta_1 x_{i1} \qquad\qquad\qquad\qquad + \varepsilon_i$ | |
| 6 | $y_i = \beta_0 \qquad\qquad + \beta_2 x_{i2} \qquad\qquad + \varepsilon_i$ | |
| 7 | $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \qquad\qquad + \varepsilon_i$ | |
| 8 | $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$ | |

*Analysis plan:*

First, test the interaction, $H_0$: $\beta_3 = 0$, with $\boldsymbol{C} = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}$ and $\theta_0 = 0$, to compare model 8 to model 7. If significant, quit.

The interaction test asks if $y_i = \alpha_0 + \alpha_1 x_{i1} + \varepsilon_i$ is the same for all $x_{i2}$.

For $x_{i2} \neq x'_{i2}$ , the null hypothesis corresponds to parallel lines with distinct intercepts (for $\beta_3 = 0$).

If $\beta_3$ is significant, both variables are important, even if $\beta_1$ and/or $\beta_2$ is not significant. If $\beta_3$ is significant, retain both variables, even if $\beta_1$ and/or $\beta_2$ is not significant.

Always include *main effects* (here, $x_1$ and $x_2$) in the model when an interaction term between them is also in the model.

## Interpreting Interaction Terms

With an interaction term in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i,$$

it is difficult to interpret $\beta_1$ and $\beta_2$ in isolation. However, we may express the effect of a one-unit increase in $x_1$ with $x_2$ held constant as

$$\begin{aligned} E(y_i|x_{i1}+1, x_{i2}) \quad - \quad & E(y_i|x_{i1}, x_{i2}) = \\ & \beta_0 + \beta_1(x_{i1}+1) + \beta_2 x_{i2} \\ & +\beta_3(x_{i1}+1)x_{i2} \\ & -(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}) \\ = \quad & \beta_1 + \beta_3 x_{i2}. \end{aligned}$$

Similarly, we can find the effect of a one-unit increase in $x_2$ with $x_1$ held constant.

## Next: Correlation

*Reading Assignment:*

- Muller and Fetterman, Chapter 6: "Correlations" (Required)

# Lecture 9: Correlations

*Reading Assignment:*

- Muller and Fetterman, Chapter 6: "Correlations"

For two random variables $X$ and $Y$, recall that the correlation $\rho$ is defined as

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},$$

where $-1 \leq \rho \leq 1$. We estimate the population correlation, $\rho$, using *Pearson's coefficient of correlation*, given by

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)}}.$$

Consider the simple linear model $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$.

- The *squared correlation coefficient*, $R^2$, measures the strength of the linear relationship between the variables $\mathbf{y}$ and $\mathbf{x}$. $R^2 \longrightarrow 1$ indicates a stronger linear relationship, and $R^2 \longrightarrow 0$ indicates a weaker linear relationship.

- The *sample correlation coefficient* $R$ is a general measure of the *linear* relationship between two random variables.
  - If $R \approx 0$, there is little evidence of a linear association between two variables. (This does not mean there is no association between the two variables.)
  - As $R \longrightarrow 1$, the linear association is more positive (i.e.,subjects with high values of $x$ will tend to have high values of $y$ as well).
  - As $R \longrightarrow -1$, the linear association is more negative (i.e., subjects with high values of $x$ will likely have low values of $y$).

We can interpret $R$ as the expected change in the response (in units of standard deviation) associated with a change of one standard deviation in the predictor of interest.

To see this, standardize the response by subtracting its mean and dividing by its standard deviation, standardize the predictor $x$ in the same way, and then regress the standardized response on the standardized predictor, omitting an intercept from the model. The resulting estimate of the regression coefficient is $R$. We verify this using the ozone data below.

```
proc corr;
var personal outdoor;
run;

proc standard data=ozone mean=0 std=1 out=ozone2;
var personal outdoor;
run;

proc glm data=ozone2;
model personal=outdoor/noint;
run;
**********************************************************************
```

```
                Pearson Correlation Coefficients, N = 64
                       Prob > |r| under H0: Rho=0

                                personal       outdoor
                 personal       1.00000       0.39916
                                                0.0011

                  outdoor        0.39916       1.00000
                                  0.0011

                          The GLM Procedure

Dependent Variable: personal

                                        Sum of
Source                     DF          Squares    Mean Square   F Value
Model                       1      10.03745850    10.03745850     11.94
Error                      63      52.96254150     0.84067526
Uncorrected Total          64      63.00000000

                     Source                  Pr > F
                     Model                   0.0010
                     Error
                     Uncorrected Total
```

```
        R-Square     Coeff Var      Root MSE    personal Mean
       0.159325    -2.8725E17      0.916883        -0.000000

NOTE: No intercept term is used: R-square is not corrected for the
      mean.

Source                     DF      Type I SS    Mean Square   F Value
outdoor                     1    10.03745850    10.03745850     11.94

                    Source                 Pr > F
                   outdoor                0.0010

Source                     DF    Type III SS    Mean Square   F Value
outdoor                     1    10.03745850    10.03745850     11.94

                    Source                 Pr > F
                   outdoor                0.0010

                                     Standard
  Parameter         Estimate            Error    t Value    Pr > |t|
  outdoor       0.3991550301       0.11551646       3.46      0.0010
```

## Example: Correlation Matrix for Ozone Data

The correlation matrix for the ozone data is provided below.

```
The CORR Procedure

         4  Variables:    personal outdoor  home     time_out

             Pearson Correlation Coefficients, N = 64

              personal       outdoor       home      time_out
 personal      1.00000       0.39916       0.53000       0.19452
 outdoor       0.39916       1.00000       0.55582       0.07818
 home          0.53000       0.55582       1.00000      -0.00714
 time_out      0.19452       0.07818      -0.00714       1.00000

             Pearson Correlation Coefficients, N = 64
```

How do we interpret these values?

These correlations may not give us exactly the information we want. Suppose that based on these correlations, we decide to perform a regression of personal exposure on home exposure. Before considering whether outdoor ozone or time spent outdoors should be added next, it would be helpful to know how associated each is with the response *after* we have controlled for home ozone.

The simple correlations do not tell us about

1. the relationship between $O_{PERSONAL}$ and $(O_{OUTDOOR}, O_{HOME},$ and $TIME_{OUT})$ as a group (multiple correlation coefficient),
2. the relationship between $O_{PERSONAL}$ and $O_{HOME}$ after controlling for $TIME_{OUT}$ (partial correlation coefficient), or
3. the relationship between $O_{PERSONAL}$ and the combined effects of $O_{OUTDOOR}$ and $O_{HOME}$ after controlling for $TIME_{OUT}$ (multiple partial correlation coefficient).

## Interpreting $\rho^2$

### Corrected $\rho^2$

Consider decomposing the variance of $Y$, namely $\sigma_y^2$. Because $\sigma_y^2 \geq 0$, it is sensible to ask what fraction (proportion) of the variance is explained by using a linear model to predict $Y$.

For a model that spans the intercept, one can define the *corrected correlation coefficient*

$$\rho_{\mathsf{c}}^2 = \frac{\sigma^2(\beta_0) - \sigma^2(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ \ldots + \beta_{p-1} X_{p-1})}{\sigma^2(\beta_0)}$$

$$\widehat{\rho}_{\mathsf{c}}^2 = R_{\mathsf{c}}^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE}$$

$$= \frac{CSS(\text{Regression})}{CSS(\text{total})}$$

Under HILE Gauss, $\widehat{\rho}_{\mathsf{c}}^2$ is the MLE. Both $0 \leq \rho_c^2 \leq 1$ and $0 \leq R_{\mathsf{c}}^2 \leq 1$.

$R^2_{\text{c}}$ equals the squared univariate correlation between $Y$ and $\widehat{Y}$ and is provided for the ozone data below.

## Example: $R^2_c$ for ozone data

Verify that the value labeled "R-Square" on the SAS output is the corrected $R^2_c$. How do we interpret this value?

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
*************************************************************************

                                        Sum of
Source                      DF         Squares    Mean Square   F Value
Model                        3      5034.90667     1678.30222      9.92
Error                       60     10148.19129      169.13652
Corrected Total             63     15183.09796

                   Source                  Pr > F
                   Model                   <.0001

        R-Square     Coeff Var      Root MSE    personal Mean
       0.331613      55.23389      13.00525         23.54578
```

## "Uncorrected" $\rho^2$

The value of uncorrected correlations lies in evaluating models without an intercept and in evaluating the value of an intercept. We follow the default of considering only corrected correlations. Note that the discussion of partial correlations generalizes to uncorrected correlations with, as always, special attention paid to the presence or absence of the intercept.

## Generalizing the Application of $\rho^2$

All GLH tests correspond to comparing two nested models, and all GLH tests correspond to comparing two correlations. We will show that the standard $F$ test (with the SSE estimated from the larger model and not the full model in this case) can be written in terms of the $R^2$ from the models being compared.

$$
\begin{aligned}
F_{obs} &= \frac{\frac{[SSE(\text{smaller}) - SSE(\text{larger})]}{[dfE(\text{smaller}) - dfE(\text{larger})]}}{\frac{SSE(\text{larger})}{dfE(\text{larger})}} \\
&= \frac{\frac{[CSS(\text{larger}) - CSS(\text{smaller})]}{[dfE(\text{smaller}) - dfE(\text{larger})]}}{\frac{SSE(\text{larger})}{dfE(\text{larger})}} = \frac{\frac{[R^2(\text{larger}) - R^2(\text{smaller})]}{a}}{\frac{[1 - R^2(\text{larger})]}{dfE(\text{larger})}}.
\end{aligned}
$$

We assume the smaller model contains $\{X_1, \ldots X_{g_1}\}$ and the larger model differs by the addition of $\{X_{g_1+1}, \ldots X_{p-1}\}$.

Let $\Delta R^2 = R^2(\text{larger}) - R^2(\text{smaller})$, with

$$0 \leq \Delta R^2 \leq 1.$$

If all variables are scaled to have unit variance, $\Delta R^2$ equals the covariance (with $Y$) of a group of predictors, adjusted for predictors already in the model. Examining that relationship leads to the study of partial correlations.

## Correlation Formulae

Let $\boldsymbol{v}_j = \{v_{ij}\}$ indicate an $n \times 1$ vector of i.i.d.observations.

- Compute the mean as $\overline{v}_j = \mathbf{J}_n' \boldsymbol{v}_j / n$.
- Compute the sample variance as

$$s_j^2 = \sum_{i=1}^{n} \left(v_{ij} - \overline{v}_j\right)^2 / n = \boldsymbol{v}_j' \boldsymbol{v}_j / n - \overline{v}_j^2,$$

  with $s_j^2 \left(\frac{n}{(n-1)}\right)$ an unbiased estimator of the population variance.

- Compute the sample covariance for two vectors $j$ and $j'$ as

$$c_{jj'} = \sum_{i=1}^{n} \left(v_{ij} - \overline{v}_j\right)\left(v_{ij'} - \overline{v}_{j'}\right) / n = \boldsymbol{v}_j' \boldsymbol{v}_{j'} / n - \overline{v}_j \overline{v}_{j'} \ ,$$

  with $c_{jj'} \left(\frac{n}{(n-1)}\right)$ an unbiased estimator of the population covariance.

- Compute the sample correlation coefficient

$$r(v_{ij}, v_{ij'}) = \frac{c_{jj'}}{s_j s_{j'}} \, .$$

## Partial Correlation

Partial correlations involve a predictor, $X$, a response, $Y$, and nuisance variables, $\mathbf{Z}$.

*Partial correlations* describe the strength of the linear relationship between two variables, $Y$ and $X$, after controlling for the effects of other variables $\mathbf{Z}$. When multiple $X$ and $Z$ variables are involved, we call the partial correlation a *multiple partial correlation*. The *order* of a partial correlation depends on the number of variables for which we control: *first-order* partials control for $\mathbf{Z}_{(n\times 1)}$, *second-order* partials control for $\mathbf{Z}_{(n\times 2)}$, and $p^{th}$*-order* partials control for $\mathbf{Z}_{(n\times p)}$.

**Example:** Why do we want to consider a partial correlation? Consider computing the correlation between hair length and height in the general population.

Consider the following table of partial correlations for the ozone data.

| *Order* | *Controlling Variables* | *Form of Correlation* | *Correlation Estimate* |
|---|---|---|---|
| 0 | | $r_{PERS,HOME}$ | 0.53 |
| 0 | | $r_{PERS,OUT}$ | 0.40 |
| 0 | | $r_{PERS,TIME}$ | 0.19 |
| 1 | HOME | $r_{PERS,OUT\|HOME}$ | 0.15 |
| 1 | HOME | $r_{PERS,TIME\|HOME}$ | 0.23 |
| 1 | OUT | $r_{PERS,HOME\|OUT}$ | 0.40 |
| 1 | OUT | $r_{PERS,TIME\|OUT}$ | 0.18 |
| 1 | TIME | $r_{PERS,HOME\|TIME}$ | 0.54 |
| 1 | TIME | $r_{PERS,OUT\|TIME}$ | 0.39 |
| 2 | HOME,OUT | $r_{PERS,TIME\|HOME,OUT}$ | 0.22 |
| 2 | HOME,TIME | $r_{PERS,OUT\|HOME,TIME}$ | 0.13 |
| 2 | OUT,TIME | $r_{PERS,HOME\|OUT,TIME}$ | 0.42 |

- Which variable has the greatest linear association with personal exposure?
- After that variable, what is the next most important variable?
- How do we tell whether the linear association is "enough" to warrant inclusion in a statistical model? When do we stop adding variables?

At some point, we need to decide whether a particular partial correlation coefficient is significantly different from zero. Previously, we used F tests to determine whether adding a variable to a regression model was worthwhile, given that other variables were in the model. This type of test is also called a *partial F test*. This test is exactly equivalent to a test of significance for the corresponding partial correlation coefficient.

That is, a test of whether the population partial correlation coefficient $\rho_{PERS,TIME|HOME}$ is equal to 0 is exactly equivalent to a test of $H_0 : \beta_{TIME} = 0$ given that home exposure is already in the model.

## Insight of Partial Correlation

Consider fitting two models, $\mathbf{y} = \mathbf{z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbf{x} = \mathbf{z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, to create $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ as well as predicted values $\widehat{\boldsymbol{y}} = \mathbf{z}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{x}} = \mathbf{z}\widehat{\boldsymbol{\gamma}}$.

Let $\widehat{\varepsilon}_{y_i} = y_i - \widehat{y}_i$ and $\widehat{\varepsilon}_{x_i} = x_i - \widehat{x}_i$.

Describe $r(\widehat{\varepsilon}_{y_i}, \widehat{\varepsilon}_{x_i})$, the correlation between these residuals, as the sample value of the (full) *partial correlation* between $Y$ and $X$, controlling both for $\mathbf{Z}$. We use the notation $r_{YX|Z}$ for this partial correlation.

The partial correlation describes the simple linear correlation between $Y$ and $X$ after removing the linear effects of $\mathbf{Z}$.

## Adjusting for Confounders

A very important application of partial F tests concerns controlling for confounders. Suppose that we have one main study variable of interest, $\mathbf{x}$, and $q$ control variables (or confounders) $\mathbf{z}_1, \ldots, \mathbf{z}_q$. We can evaluate the effect of $\mathbf{x}$ on the outcome of interest, controlling for the potential confounders, $\mathbf{z}$, using the model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{z}_1 + \ldots + \beta_q \mathbf{z}_q + \beta_{q+1} \mathbf{x} + \boldsymbol{\varepsilon}.$$

We then test the hypothesis $H_0 : \beta_{q+1} = 0$, which is equivalent to testing whether the population partial correlation between $\mathbf{y}$ and $\mathbf{x}$, adjusting for the confounders $\mathbf{Z}$, is zero.

When we are interested in several variables in $\mathbf{X}$, we must determine which of these are important, and we may need to rank order them by their relative importance. We will discuss this strategy further when we talk about model selection.

## Semi-Partial Correlation

We consider semi-partial correlations when we know that only one of $X$ or $Y$ is affected by the nuisance variables $Z$. For example, in a randomized clinical trial, the treatment variable $X$ should be unaffected by nuisance factors $Z$ (such as age or tumor type), although these nuisance factors may have an effect on the outcome.

Describe $r(\widehat{\varepsilon}_{y_i}, x_i) = r_{X(Y|Z)}$ as the semi-partial correlation between $Y$ and $X$, controlling only $Y$ for $\mathbf{Z}$.

Also define $r(y_i, \widehat{\varepsilon}_{x_i}) = r_{Y(X|Z)}$ as the semi-partial correlation between $Y$ and $X$, controlling $X$ for $\mathbf{Z}$.

## Example: Semi-Partial Correlations in Ozone Data

Suppose we believe that while time spent outdoors may be related to personal ozone exposure, time spent outdoors and home ozone concentration are unlikely to be related (r=-0.007). In this case, when computing the correlation between personal ozone exposure and home ozone concentrations, we may wish to correct personal ozone exposure (but not home ozone concentrations) for the effect of time spent outdoors.

First, we will compute the partial correlation between personal ozone exposure and home ozone concentration, controlling both variables for the effect of time spent outdoors.

```
proc corr data=ozone nosimple /* noprob */;
var personal home;
partial time_out;
run;
*************************************************************************
```

```
          Pearson Partial Correlation Coefficients, N = 64
                Prob > |r| under H0: Partial Rho=0

                                              personal          home

 personal                                      1.00000       0.54175
 Personal Ozone Exposure (ppb)                                  <.0001

 home                                          0.54175       1.00000
 Home Indoor Ozone Concentration (ppb)          <.0001
```

So the partial correlation between personal ozone exposure and home ozone concentration, controlling both variables for the effect of time spent outdoors is 0.54.

Next, we control only personal ozone exposure (and not home ozone concentration) for the effect of time spent outdoors.

```
proc corr data=semipart nosimple noprob;
var home e_p;
run;
*********************************************************************
                    Pearson Correlation Coefficients, N = 64

                                                   home           e_p
  home                                          1.00000       0.54173
  Home Indoor Ozone Concentration (ppb)
  e_p                                           0.54173       1.00000
```

The partial and semi-partial correlation coefficients are virtually identical. The simple correlation between home ozone concentration and personal ozone exposure is also similar at 0.53, suggesting that the linear influence of time spent outdoors on personal ozone exposure and/or on home ozone concentrations has little effect on the correlation between personal ozone exposure and home ozone exposure.

## Choosing the Proper Correlation Coefficient

Considering the proper correlation coefficient to describe the relationship among variables $Y$ and $X$ along with nuisance (or confounding) variables $Z$ depends on the nature of their relationship. The table below can be used to determine which correlation is appropriate.

| *Nuisance Relationship* | *Preferred Correlation* |
|---|---|
| Neither $X$ nor $Y$ affected by $Z$ | $r_{xy}$ |
| Both $X$ and $Y$ affected by $Z$ | $r_{yx\|z}$ |
| Only $X$ affected by $Z$ | $r_{y(x\|z)}$ |
| Only $Y$ affected by $Z$ | $r_{x(y\|z)}$ |

## Computing Partial Correlations

Consider the following correlation matrix for three variables $X$, $Y$, and $Z$.

$$\boldsymbol{R} = \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{bmatrix} 1 & r_{xy} & r_{xz} \\ r_{yx} & 1 & r_{yz} \\ r_{zx} & r_{zy} & 1 \end{bmatrix} \end{array} \begin{array}{c} \\ X \\ Y \\ Z \end{array}$$

Note that $r_{yx} = r_{xy}$.

A full partial correlation may be computed as:

$$r_{yx|z} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1 - r_{yz}^2)(1 - r_{xz}^2)}} \, .$$

In the same setting, semi-partials may be computed as:

$$r_{y(x|z)} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1)(1 - r_{xz}^2)}}$$

$$r_{x(y|z)} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{(1 - r_{yz}^2)(1)}} \, .$$

The $(1)$ term represents the variance of a standardized variable.

## Summary: Partial Correlation Coefficient

- The partial correlation $r_{YX|Z_1,\ldots,Z_q}$ measures the strength of the linear relationship between $X$ and $Y$ while controlling for $\mathbf{Z}$.
- The square of the partial correlation $r_{YX|Z_1,\ldots,Z_q}$ measures the proportion of the sum of squares for error in a model containing only $\mathbf{Z}$ that is accounted for by the addition of $X$ to a regression model already containing $\mathbf{Z}$.
- The partial $F$ statistic is used to test $H_0 : \rho_{YX|Z_1,\ldots,Z_q} = 0$.
- The partial correlation $r_{YX|Z}$ can be defined as the correlation of the residuals of the straight-line regressions of $Y$ on $Z$ and of $X$ on $Z$.
- Multiple partial correlations are tested using group-wise partial $F$ tests.

## Next: Assumption Diagnostics

*Reading Assignment:*

- Muller and Fetterman, Chapter 7: "GLM Assumption Diagnostics"
- Weisberg, Chapter 9: "Regression Diagnostics"

# Lecture 10: Assumption Diagnostics

*Reading Assignment:*

- Muller and Fetterman, Chapter 7: "GLM Assumption Diagnostics" (Required)

Wait, what are those assumptions agagin?

Homogeneity, Independence, Linearity, Existence, and Gaussian distribution.

Checking model assumptions focuses on analysis of residuals and outliers. Ethically, one must avoid allowing exploratory analysis (data snooping) to bias confirmatory analysis (inflate $\alpha$). It is, however, acceptable to examine outliers or to transform variables to get the best test of the Gaussian distribution of residuals.

## The First Step: Get to Know Your Data

Checking the Basics

- Determine the sampling unit (child, family, mouse, basketball team)
- Investigate the data collection procedure (counting number of vegetable servings consumed daily, measuring tumor size, measuring body mass index (BMI))
- Obtain units of measurement for all variables (number, cubic centimeters, kilograms per square meter)
- Obtain plausible range of values (0-10 servings, 0-1000 cubic centimeters, 17-40 kg per square meter)
- Obtain typical values for all variables (2 servings, 25 cubic cm, 25 kg per square meter)

Examining Individual Values

- Print the first 50 or so observations (PROC PRINT DATA=OZONE (OBS=50);)
- Calculate the minimum and maximum values for all numeric variables (PROC MEANS MIN MAX;)
- Generate frequency tables for character variables (PROC FREQ)

Summarizing the Data

- Report descriptive statistics (mean, median, quartiles, largest and smallest few observations - PROC UNIVARIATE)
- Generate appropriate graphical displays (histograms, stem and leaf plots, etc.using PROC UNIVARIATE)
- Look for patterns and unusual values

Check correlations, covariances, plot $X_j$ versus $Y$, $X_j$ versus $X_{j'}$

## Example: Basic Diagnostics for the Ozone Data

The following SAS code is used to provide basic diagnostics.

```
proc print data=ozone (OBS=50);
run;

proc means min max data=ozone;
run;

proc univariate plot normal data=ozone;
var personal;
run;

proc capability data=ozone;
qqplot personal outdoor home time_out;
histogram time_out;
run;

proc corr noprob data=ozone;
var personal outdoor home time_out;
run;
```

The art in reporting these diagnostics lies in doing enough to provide necessary information with sufficiently few pages to be read (if you are

bored with your report, certainly everyone else is!). See Example 7.1 in the textbook for a nice illustration.

Selected diagnostics from the ozone data are presented below.

1. Basics: the sampling units in the ozone data are young children. Personal ozone exposures were measured using a portable monitor attached to the child's clothing, outdoor ozone exposures were measured at EPA monitoring stations, indoor (home) exposures were measured using monitors in each child's home, and the proportion of time spent outdoors was reported by a parent. Ozone exposures are measured in parts per billion. The plausible range of time spent outdoors would be between 0 and 1.

2. Individual values: in the list of data, one subject spent 90% of time outdoors, leading the biostatistician to make a note to ask the investigator if this variable is "proportion of play time spent outdoors" because 90% of one day would be 21.6 hours.

3. Data Summary: personal ozone exposures range from 0.46 to

67.18 ppb, with one outlier on the boxplot (67.18 ppb, which is not disturbing in a sample of 64 observations) and a mean and median in the range of 22-24 ppb. This variable does not look perfectly Gaussian, which is not disturbing as the Gaussian assumption applies not to $Y$ unconditionally but to $Y$ conditional on the covariates. (We should describe the covariates in a similar manner, noting that we need not assume they follow Gaussian distributions.)

4. Further details: personal ozone exposure appears to be most highly correlated with home exposure, and we do not see evidence of collinearity of covariates from the correlations.

## Violations of Independence

The independence assumption applies to all statistical tests we have discussed thus far. The consequences of violating the independence assumption range from finding spurious relationships to missing significant ones.

Common violations of the independence assumption include

- repeated measures
  - weight loss of individuals is monitored over time
  - CD-4 counts of AIDS patients are recorded over the course of the disease
- nested data
  - schools are randomly assigned different physical education programs with student fitness as the outcome

- family data
  - researchers at NIEHS are studying sisters and daughters of women with breast cancer
  - pregnant rats are exposed to toxins and litters are studied for malformations

In general, independence must be evaluated by careful consideration of the study design. Warning signs include "too many" data points (often lab scientists run multiple assays on tissues from one specimen, for example) or "strange" results.

## Residual Analysis

Recall that the raw residuals are estimates of errors for the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

The HILE Gauss assumptions refer to errors ($\boldsymbol{\varepsilon}$), which correspond to $Y|X$ and not $Y$ unconditionally.

Thus to evaluate the assumptions we study the residuals $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, with $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

## Example: Residuals

Suppose we plan to sample undergraduate males at UNC and measure their lung function.

What might the density (histogram) of their lung function look like?

(What fraction smoke?) $Y$ may be bimodal (if so, certainly not normal!). However, errors, if we include the amount smoked in our model, may be Gaussian.

## Properties of Residuals

Recall that the raw residuals, which are differences between the data points and fitted values, follow a singular normal distribution

$$\widehat{\boldsymbol{\varepsilon}} \sim \mathcal{SN}_n\{\mathbf{0}, \sigma^2[\boldsymbol{I} - \boldsymbol{H}]\}$$

and that we can estimate $\sigma^2$ using the residuals as follows:

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2/(n-r).$$

The covariance among residuals, $\sigma^2[\boldsymbol{I} - \boldsymbol{H}]$, implies non-zero correlation and non-independence among residuals, though such correlations are modest in many settings.

If the model spans an intercept, then $\overline{\widehat{\varepsilon}} = \sum_{i=1}^{n} \widehat{\varepsilon}_i/n = 0$. In this case, it is clear that residuals are *not* independent!

## Standardized Residuals

Because $\text{Var}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_i)$, the residuals may have different variances because the variance of the residuals depends on the pattern of covariate values.

It would be easier to use the residuals to evaluate model assumptions if they all had the same variance. For this reason, we construct *standardized residuals*

$$r_i = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2(1 - h_i)}}.$$

The standardized residuals (also called internally studentized residuals) will follow a Gaussian distribution with mean 0 and variance 1 in large samples. Because of this property, one rule of thumb is to examine any observations with standardized residuals $> 2$ in absolute value in further detail (the justification is that 95% of the mass of the Gaussian distribution lies between -1.96 and 1.96, so that studentized residuals with absolute values greater than 2 represent observations in the

extreme tails of the distribution). Standardizing residuals thus helps us to assess their magnitude relative to the precision of the estimated regression analysis. Although these residuals are standardized (and thus have mean zero and variance one), they do depend on a variance estimate that may be affected by outliers. Because of this, we may miss those outliers in a residual analysis.

**The jack-knifed or studentized residuals** provide a solution to this problem based on more robust variance estimates. Recall that the studentized residuals (our preference) are defined as

$$r_{(-i)} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2_{(-i)}(1-h_i)}} \sim T(n-r-1),$$

where the jack-knifed estimate of $\sigma^2$ is more robust to outliers than the usual estimate. These residuals are also called externally studentized residuals.

Note that a sample set of standardized or studentized residuals have approximate mean zero and approximate variance one. As $(n-r)$ becomes large, $T \rightarrow$Gaussian. Studentized residuals can still be thought of in "standard deviation" units so that the usual cutoff of $2$ is approximately valid.

## Evaluating Assumptions with (or Without) Residuals

- Homogeneity: violations seen in the pattern of residuals.
- Independence: assessed through logic of sampling scheme.
- Linearity: examine pattern of residuals.
- Existence: (finite sample...).
- Gaussian distribution: distributional assessment involves box plot of residuals, histogram of residuals, and test of Gaussian distribution of residuals. (The discrepancy between $T$ and Gaussian random variables somewhat inflates the probability of rejecting the null...why?)

Sometimes, Gaussian errors, homogeneity, and linearity come and go as a package.

## Evaluating Heterogeneity and Linearity

Plotting $\{r_{(-i)}\}$ vs. $\{\widehat{y}_i\}$ in an *R/P plot* provides the most useful diagnostic because predicted values capture all the information in predictors that is available as a linear combination of them. The R/P plot allows us to assess both heterogeneity and linearity. The R/P plot should resemble a rectangular cloud with no obvious trends or pattern. When examining an R/P plot, look for the following indicators of problems:

- any trend, such as a tendency for negative residuals for small predicted values and positive residuals for large predicted values, which may indicate non-linearity
- non-constant spread of the residuals, such as a tendency for tightly clustered residuals for small predicted values and widely dispersed residuals for large predicted values, which may indicate heteroscedasticity.

One issue in multiple regression is the detection of *multivariate* outliers. The R/P plot replaces a vector of predictors with one univariate predicted value. This is just one way to move from the multivariate predictor space to a univariate space.

A *probability plot* or *Q-Q* plot graphs the ordered residuals versus the expected order statistics of the standard normal distribution. To create a Q-Q plot, we order the residuals from smallest to largest, letting $r_{(1)}$ be the smallest, $r_{(i)}$ the $i^{th}$ smallest, and $r_{(n)}$ be the largest. As you may recall from BIOS 550/660, $r_{(i)}$ is called an *order statistic*.

Next, imagine taking a sample of size $n$ from a standard normal distribution and ordering it to obtain the order statistics $z_{(i)}$, $i = 1, \ldots, n$. Using results from probability theory, we can obtain the expected values of these order statistics. (For example, we expect the median to be 0, the $5^{th}$ percentile to be -1.96, and the $95^{th}$ percentile to be 1.96.)

If our observations come from a Gaussian distribution, we expect our observed order statistics to be pretty close to the expected order statistics from the standard normal distribution. Specifically, if we plot the observed order statistics against the expected ones, we expect to

see a straight line through the origin with unit slope. When this is not the case, the observed order statistics deviate from what we would expect from a standard normal distribution, and the Gaussian errors assumption may be violated.

For example, if the residuals are from a distribution with heavy tails, we expect to see values consistently above a straight line through the origin with unit slope in the upper tail and consistently below the line in the lower tail. If the residuals are from a distribution with light tails, then we expect to see values consistently below the line in the upper tail and above the line in the lower tail.

A curve that is concave upwards indicates positively-skewed data, and a curve that is concave downwards indicates negatively-skewed data.

Standard Normal Data
Frequency
0
500
1500
−4
−2
0
2
4
xa
Standard Normal Data
Sample Quantiles
−4
−2
0
2
4
Theoretical Quantiles
Heavy−Tailed Data
Frequency
0
1000
2500
−10
−5
0
5
10
xb
Heavy−Tailed Data
Sample Quantiles
−10
−5
0
5
10
Theoretical Quantiles
Light−Tailed Data
Frequency
0
200
600
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
xc
Light−Tailed Data
Sample Quantiles
−1.5
0.0
1.0
Theoretical Quantiles

Positively−Skewed Data
Frequency
0
500
1500
−2
0
2
4
6
xd
Positively−Skewed Data
Sample Quantiles
−2
0
2
4
6
−4
−2
0
2
4
Theoretical Quantiles
Negatively−Skewed Data
Frequency
0
500
1500
−6
−4
−2
0
2
xe
Negatively−Skewed Data
Sample Quantiles
−6
−4
−2
0
2
−4
−2
0
2
4
Theoretical Quantiles

The following code creates an R/P plot as well as a Q-Q plot of the studentized residuals.

```
proc reg data=ozone;
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted h=leverage;
run;

proc plot data=out;
plot studresid*predicted/vref=0;
run;

proc capability data=out;
qqplot studresid;
run;
```

Now we consider a variety of example $R/P$ plots in order to see what types of plots are acceptable and what types of plots are cause for concern.

Figure 1: Acceptable residuals: simulated $x_i \sim \mathcal{N}(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + \varepsilon_i$ and fit valid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Figure 2: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + x_i\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Figure 3: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + (1 - x_i)\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Figure 4: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + 2 \mid 0.5 - x_i \mid \varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Figure 5: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + (1 - \mid 0.5 - x_i \mid)\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Figure 6: Unacceptable residuals (violation of linearity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i^2 + 0.05\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Some analysts plot the response versus each individual predictor, a tactic that ignores the impact of other predictors. A better idea is to create the *partial plot* of $\{\mathbf{y}\}$ versus $\{\mathbf{x}_j|\{\mathbf{x}_1, \ldots \mathbf{x}_{j-1}, \mathbf{x}_{j+1} \ldots, \mathbf{x}_p\}\}$, the residuals from a regression with $\mathbf{x}_j$ as response and the other covariates as the predictors. The univariate correlation in the plot is the semi-partial correlation

$$r(\mathbf{y}, \mathbf{x}_j \mid \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p\}).$$

These plots are most useful for identifying culprits in known problems and not for detecting any new problems.

## Evaluating Gaussian Distribution

SAS PROC UNIVARIATE provides the following information:

- a Q-Q ("normal") plot (use PROC CAPABILITY to get a nicer one),
- a box and whisker plot,
- a stem and leaf plot (a very basic histogram),
- moments (mean, variance, skewness, kurtosis),
- the largest and smallest 5 values with ID's, and
- quartiles.

## Example: Ozone Data Residuals

The following code is used to evaluate the residuals for the ozone data from the model

$$O_{PERSONAL} = \beta_0 + \beta_1 \, O_{OUTDOOR} + \beta_2 \, O_{HOME} + \beta_3 \, TIME_{OUT} + \varepsilon.$$

```
proc reg data=ozone;
```

```
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted;
run;

proc univariate plot normal data=out;
id subject;
var studresid;

proc capability data=out;
qqplot studresid;
run;

proc capability data=out;
histogram studresid/kernel;
run;

*********************************************************************
                      The UNIVARIATE Procedure
   Variable:  studresid  (Studentized Residual without Current Obs)

                                Moments

   N                          64    Sum Weights                 64
   Mean               0.00618379    Sum Observations    0.39576231
   Std Deviation       1.0298702    Variance            1.06063263
   Skewness           0.90708362    Kurtosis              0.893602
```

```
Uncorrected SS     66.8223027    Corrected SS        66.8198554
Coeff Variation    16654.3631    Std Error Mean      0.12873377

                   Basic Statistical Measures

         Location                     Variability

      Mean      0.00618     Std Deviation            1.02987
      Median   -0.33120     Variance                 1.06063
      Mode        .         Range                    5.12982
                            Interquartile Range      1.02182

                   Tests for Location: Mu0=0

         Test           -Statistic-    -----p Value------

         Student's t    t  0.048035    Pr > |t|    0.9618
         Sign           M        -4    Pr >= |M|   0.3817
         Signed Rank    S      -133    Pr >= |S|   0.3779

                      Tests for Normality

      Test                  --Statistic---    -----p Value------

      Shapiro-Wilk          W     0.935601    Pr < W      0.0024
      Kolmogorov-Smirnov    D     0.148292    Pr > D     <0.0100
```

```
Cramer-von Mises      W-Sq  0.279391    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  1.565583    Pr > A-Sq  <0.0050
```

We see that all the normality tests agree that the residuals do not appear to be normal. (P-value shopping is not advised – it is best to pick one test before doing the analysis.) The differences among the tests are described below. For all tests, the null hypothesis is $H_0$ : the data are normally distributed.

- Shapiro-Wilks: roughly, a measure of the straightness of a Q-Q plot (appropriate for small sample sizes)
- Kolmogorov-Smirnov: largest discrepancy between the empirical cdf and the estimated hypothesized one (based on observed mean and variance)
- Cramer-von Mises: considers squared difference between empirical and estimated cdf
- Anderson-Darling: considers a weighted squared difference between empirical and estimated cdf

We can look at the quantiles, boxplot, and histogram to check whether residuals are skewed.

```
Quantiles (Definition 5)

Quantile       Estimate

100% Max       3.054968
99%            3.054968
95%            2.155071
90%            1.483757
75% Q3         0.388099
50% Median    -0.331204
25% Q1        -0.633717
10%           -0.907937
5%            -1.151776
1%            -2.074857
0% Min        -2.074857
```

We see some evidence of skewness in the quantiles.

In addition, we can look at the most extreme residuals to gauge whether any outliers deserve further attention.

```
Extreme Observations
```

```
-----------Lowest----------        ----------Highest----------

   Value   subject      Obs           Value   subject      Obs

-2.07486        25       25         1.70690        49       49
-1.78561        36       36         2.15507        64       64
-1.51369         4        4         2.33176        17       17
-1.15178         2        2         2.67945        39       39
-1.08553        20       20         3.05497        26       26
```

There do seem to be some outliers on the right.

```
Stem Leaf                     #             Boxplot
  18
  16 1                        1                |
  14 87                       2                |
  12 09                       2                |
  10 65                       2                |
   8 1                        1                |
   6 53                       2                |
   4 8                        1                |
   2 358135889                9             +-----+
   0 0360                     4             |  +  |
  -0 415                      3             |     |
  -2 998                      3             *-----*
  -4 9863385441              10             |     |
  -6 729852110                9             +-----+
  -8 19520                    5                |
 -10 596                      3                |
 -12                                           |
 -14 1                        1                |
 -16 9                        1                |
 -18                                           |
 -20 7                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-1
```

Studentized Residual

4
3
2
1
0
-1
-2
-3
Studresid
-3
-2
-1
0
1
2
3
Normal Quantiles

From this output, we see that the data appear to be skewed to the right a little bit, and there is some peakedness in the distribution. Again, we see several outliers in the right tail of the distribution. A *kernel density estimate* is superimposed on the histogram in PROC CAPABILITY in order to facilitate interpretation.

## Evaluating Extreme Residuals

One residual must be the most extreme, but too many extreme residuals (or too extreme residuals) are evidence of poor model fit.

Testing residuals (the null hypothesis is $H_0$: $E[r_{(-i)}] = 0$) leads to a multiple comparisons issue: we do $n$ tests (one for each residual), so we should use a Bonferroni correction (use $\alpha/n$ instead of $\alpha$ and the $\alpha/(2n)$ critical value for $T$).

Any positive correlation among residuals makes the correction conservative.

If $(n - r) \to \infty$ then $\sigma^2(\boldsymbol{I} - \boldsymbol{H}) \to \sigma^2 \boldsymbol{I}$, implying that correlations among residuals converge to zero for large samples.

## Example: Extreme Residuals in Ozone Data

Based on previous analyses, do any residuals appear to be extreme? If so, which ones? What do you conclude based on a test of the hypothesis that any extreme residuals come from a population with mean 0?

## Outliers

An *outlier* is a value (of a predictor or a response) much larger in absolute value than next nearest value. On a box plot, outliers are often marked using 0 or $*$.

Least squares estimation is rather sensitive to outliers.

An outlier may be an anomaly or merely a chance event (e.g.,a child reports 27 vegetable servings per day, or height is reported as 60 feet rather than 60 inches).

Automatically discarding extreme observations is WRONG! Exceptions: verifiable instrument malfunction, recording error.

Consider whether an outlier is plausible, implausible, or impossible.

- A woman weighs 250 pounds, 400 pounds, or 1000 pounds...
- Body temperature of 36.8 (centigrade or Farenheit? average

person, heart surgery patient, or crocodile?)

Extreme residuals may indicate an anomaly, while extreme values of $Y$ or $X_j$ need not be important.

**Leverage**

A *leverage* value depends only on $\mathbf{X}$ and measures how extreme the $i$th observation is in terms of the predictor space. An observation with high leverage has the potential to have great influence on the model fit.

The hat matrix, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, has $i$th diagonal element $h_i$, the *leverage* for the $i$th observation.

$h_i = \boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i'$ ,with $\boldsymbol{X}_i$ $1 \times p$, the $i$th row of $\boldsymbol{X}$

For $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $s_x^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n$,

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{n s_x^2} = \left[1 + \frac{(x_i - \overline{x})^2}{s_x^2}\right]/n\,.$$

With an intercept and all $p-1$ predictors mean zero and uncorrelated,

$$h_i = \frac{1}{n} + \sum_{j=1}^{p-1} \frac{x_{ij}^2}{n s_{x_j}^2} = \left[1 + \sum_{j=1}^{p-1} \frac{x_{ij}^2}{s_{x_j}^2}\right]/n\,.$$

We can prove $\sum_{i=1}^n h_i = r$, the rank of $\boldsymbol{X}$, so full rank $\boldsymbol{X}$ implies $\overline{h} = \sum_{i=1}^n h_i/n = p/n$.

A handy rule of thumb is to examine further any subjects with $h_i > \frac{2p}{n}$, which is leverage greater than twice the average value.

With an intercept and $(p-1)$ multivariate Gaussian predictors (NOT part of HILE Gauss!) with each row i.i.d., we see that a function of

the leverage values follow an $F$ distribution,

$$F_i = \frac{(h_i - 1/n)/(p-1)}{(1-h_i)/(n-p)} = F(p-1, n-p)\,.$$

Multiple testing leads to using a Bonferroni correction and using $\alpha/n$.

This $F$ statistic tests the hypothesis that the observation in question comes from a multivariate Gaussian distribution with the same mean as the other observations.

If the model does not span an intercept, then

$$F_i = \frac{h_i/p}{(1-h_i)/(n-p)} \sim F(p, n-p)\,.$$

## Example: Leverage Values for Ozone Data

The following code is used to print and test leverage values.

```
proc reg data=ozone;
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted h=leverage;
run;
proc sort data=out;
by descending leverage;
run;
data out_l;
set out (obs=10);
p=4; n=64;
F=((leverage-(1/n)/(p-1)))/((1-leverage)/(n-p));
pvalue=1-probf(F,p-1,n-p);
if pvalue <=0.05/n then BONF="*";
                   else BONF=" "; *Bonferroni correction;
label Bonf="Signif at 0.05/n?";
run;
proc print data=out_l uniform label noobs;
var subject personal outdoor home time_out leverage F pvalue Bonf;
run;
```

```
*********************************************************************
            Personal                        Home Indoor     Proportion
```

| subject | Ozone Exposure (ppb) | Outdoor Ozone Concentration (ppb) | Ozone Concentration (ppb) | of Time Spent Outdoors |
|---|---|---|---|---|
| 9 | 28.55 | 92.563 | 7.14 | 0.26 |
| 13 | 38.28 | 30.435 | 35.38 | 0.69 |
| 2 | 14.63 | 43.792 | 13.97 | 0.90 |
| 12 | 37.58 | 104.100 | 45.10 | 0.42 |
| 35 | 11.81 | 20.649 | 5.71 | 0.78 |
| 7 | 31.97 | 88.863 | 44.12 | 0.05 |
| 42 | 61.19 | 54.318 | 46.04 | 0.45 |
| 17 | 53.18 | 70.003 | 11.13 | 0.65 |
| 8 | 32.45 | 81.916 | 45.74 | 0.33 |
| 30 | 13.94 | 71.878 | 9.51 | 0.38 |

| Leverage | F | pvalue | Signif at 0.05/n? |
|---|---|---|---|
| 0.21178 | 15.7246 | .000000117 | * |
| 0.15601 | 10.7210 | .000009731 | * |
| 0.15214 | 10.3980 | .000013240 | * |
| 0.14355 | 9.6918 | .000026235 | * |
| 0.13310 | 8.8519 | .000060328 | * |
| 0.12067 | 7.8783 | .000162823 | * |

```
0.11869     7.7259    .000190717       *
0.11679     7.5798    .000222079       *
0.09578     6.0096    .001193472
0.09544     5.9852    .001225971
```

How do you interpret the results?

## Mahalanobis Distance

Assume the model spans an intercept and let $\mathbf{X}^*$ indicate the $p-1$ predictors other than the intercept (removing the intercept column from the $\mathbf{X} = \begin{bmatrix} \mathbf{J}_n & \mathbf{X}^* \end{bmatrix}$ matrix).

Define $\mathbf{C}^*$ as the $(p-1) \times (p-1)$ sample covariance matrix for the predictors in $\mathbf{X}^*$:

$$\mathbf{C}^* = \frac{\mathbf{X}^{*\prime} \left( \mathbf{I}_n - \frac{\mathbf{J}_n \mathbf{J}_n'}{n} \right) \mathbf{X}^*}{n}.$$

(Note that the sample variances of the predictors lie on the diagonal of $\mathbf{C}^*$.)

Compute the unbiased estimate of population covariances as

$$\widehat{\boldsymbol{\Sigma}^*} = \left( \frac{n}{n-1} \right) \mathbf{C}^*.$$

The *Malhalanobis distance*

$$m_i = (\boldsymbol{X}^{*\prime}_i - \overline{\boldsymbol{x}^*})' \widehat{\boldsymbol{\Sigma}^*}^{-1} (\boldsymbol{X}^{*\prime}_i - \overline{\boldsymbol{x}^*})$$

is the deviation of one observation's predictors from the center of the predictor space.

If all predictors are uncorrelated (in the sample at hand), then

$$m_i = \sum_{j=1}^{p-1} \frac{(x^*_{ij} - \overline{x^*}_j)^2}{\widehat{\sigma}^2_{x^*_j}} \, .$$

## The Essential Equivalence of Leverage and Malhalanobis Distance

Both leverage and Malhalanobis distance measure the "extremeness" of an observation in $\mathbf{X}$ space.

For any model that spans an intercept,

$$h_i = n^{-1} + (n-1)^{-1} m_i.$$

Thus only leverage or Malhalanobis distance need be examined (leverage is more commonly used).

## Influence: Cook's Distance

Cook proposed an influence measure based on the extent to which parameter estimates would change if we had deleted the $i^{th}$ observation from the sample. *Cook's distance* assesses the impact of deleting one observation from the sample.

Why is this a sensible diagnostic?

Cook's statistic measures the standardized shift in predicted values and the shift in $\widehat{\boldsymbol{\beta}}$ due to deleting the $i$th observation:

$$\begin{aligned} D_i &= \frac{(\widehat{\boldsymbol{y}}_{(-i)} - \widehat{\boldsymbol{y}})'(\widehat{\boldsymbol{y}}_{(-i)} - \widehat{\boldsymbol{y}})}{p\widehat{\sigma}^2} \\ &= \frac{(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} \\ &= r_i^2 \cdot \frac{h_i}{p(1-h_i)} \end{aligned}$$

where $r_i$ is standardized residual. As a quadratic form, $D_i \geq 0$. Distributional results are not straightforward, and there is no perfect rule of thumb for evaluating any particular $D_i$. One proposed rule of thumb is that values of $D_i$ close to 1 are indicative of excessive influence.

## DFBETAS and DFFITS

Two other common measures of influence are DFBETAS and DFFITS.

The *DFFITS* statistic is a scaled measure of the change in the predicted value of the $i^{th}$ observation if that observation were omitted from the analysis. That is,

$$DFFITS_i = \frac{\widehat{y}_i - \widehat{y}_{(-i)}}{\sqrt{\widehat{\sigma}^2_{(-i)} h_i}}.$$

A general rule of thumb is to examine further any observations with $DFFITS_i > 2\sqrt{\frac{p}{n}}$.

The *DFBETAS* statistic is a normalized measure of the effect of observations on the estimated regression coefficients. There are multiple DFBETAS statistics for each subject (one for each regression

coefficient). They are computed as

$$DFBETAS_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(-i)}}{\sqrt{\widehat{\sigma}^2_{(-i)}(\mathbf{X}'\mathbf{X})^{-1}_j}}.$$

One typically examines observations with $DFBETAS_{ij} > \frac{2}{\sqrt{n}}$.

## Concluding Comments

"One should be cautioned that deleting the most deviant observations will in all cases slightly improve, and sometimes substantially improve, the fit of the model. One must be careful not to data snoop simply in order to polish the fit of the model by discarding troublesome data points." (KKM, 1988, p201)

KKM: Kleinbaum, Kupper, Muller, and Nizam, *Applied Regression Analysis and Other Multivariable Methods*

One *must* report any deletions, and observations should be deleted only in the most extreme circumstances.

Leverage evaluates design and extremeness in predictor space, while residuals evaluate model adequacy. Cook's $D_i$ values evaluate influence (both of above).

It may be argued that a measure of adequacy of a sample is that no

observation is influential, even though it may be high in leverage and have an extreme residual.

Use $\{h_i\}$, $\{m_i\}$, $\{r^2_{(-i)}\}$, $\{\widehat{\boldsymbol{\beta}}_{(-i)}\}$, and $\{\widehat{\sigma}^2_{(-i)}\}$ to define "bad data," not $\boldsymbol{y}$ and $\boldsymbol{X}$.

## Next: Computation Diagnostics

*Reading Assignment:*

- Muller and Fetterman, Chapter 8: "GLM Computation Diagnostics"

# Lecture 11: Computation Diagnostics

*Reading Assignment:*

- Muller and Fetterman, Chapter 8: "GLM Computation Diagnostics" (Required)

Naive acceptance of computer output may lead to major errors, even though all assumptions prove valid.

The culprit is finite precision arithmetic, which is typically 7-15 decimal digits of accuracy for a single computation. Regression computations accumulate sums and cross-products so that numerical inaccuracy accumulates.

In the ideal situation, you are interested in a specific number of predictors and know the model beforehand. In many observational studies, investigators have a large number of measurements of predictors or exposures and want to evaluate them all in a model (e.g., positive anxiety (promotion at work, engagement), negative anxiety (fired, illness of family member), impact of positive anxiety, impact of negative anxiety, and potential interactions of these with two stress hormones!).

Including a lot of variables in a model often results in multicollinearity, a high degree of correlation among several predictors. This happens when too many variables have been put into the model, and a number of these variables measure similar phenomena. Collinearity does not cause a violation of the HILE Gauss assumptions. However, it

1. tends to inflate the variances of predicted values, and
2. tends to inflate the variances of parameter estimates (bad!).

## Single Variable Problems and Solutions

Disparity in location and scale (mean and variance) can create substantial inaccuracy. Examine summary statistics (mean, variance, minimum, and maximum) to find potential location or scale problems. The easiest remedy is to scale predictors to have range $\in \{-10, +10\}$, a common scientific practice.

Examples of location or scale problems:

- Consider children's birth year, ranging from 1980 to 1990. This creates a location problem (and *collinearity* with the intercept), but replacing birth year with age avoids the location problem.
- Consider children's spirometry performance measured by forced vital capacity (FVC, the maximum amount of air that can be forcibly expired). Recording values in mL may give a range of $\{1000, 4000\}$ mL, leading to very large elements in $\mathbf{X}'\mathbf{X}$. Using liters (dividing FVC by 1000) eliminates the scaling problem.

Reducing location and scale disparities never hurts and may substantially improve accuracy. In extreme cases, creating means of zero (centering) and sums of squares or variances of one (normalizing) may be necessary, though using convenient center points and scales often suffices. For example, use human body temperatures as $T - 37$ (C).

## Collinearity

Predictors in a model are *collinear* whenever the columns of $\mathbf{X}$ contain some amount of redundancy.

Mathematically, collinearity corresponds to linear dependence among columns of $\mathbf{X}$. Collinearity exists along a continuum (from nonexistent to moderate to severe to disastrous). Often, collinearity involves more than two variables.

**Classic Example: Socioeconomic Status**

Outcome: anything
Predictors: gender, race, age, education, socioeconomic status

Collinearity diagnostics identify variables containing little or no information above and beyond that in the other predictors.

Recall that the # of linearly independent columns in $\mathbf{X}$ equals $r =$rank$(\mathbf{X})$, and a full rank model has rank$(\mathbf{X}) = r = p$.

Some ANOVA (Chap. 12) designs use purposefully less than full rank $\mathbf{X}$, but for now the focus centers on unintentionally less than full rank designs (i.e.,problems).

Uncontrolled collinearity creates computational difficulties and confuses testing and interpretation.

Loosely, consider

- Rank: the number of distinct dimensions of information.
- Eigenanalysis: creating sets of regression coefficients (eigenvectors) needed to produce new variables which are uncorrelated, with successively maximum variances (eigenvalues).

Consider the columns of $\mathbf{X}$ as a collection of vectors:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_1 & \dots & \mathbf{X}_{p-1} \end{bmatrix}.$$

Here $\mathbf{X}_0 = \mathbf{J}_n$ whenever the model contains an intercept.

Recall that the columns of a matrix $\mathbf{X}$ are linearly dependent if there exists a vector $\boldsymbol{\delta} \neq 0$ such that $\mathbf{X}\boldsymbol{\delta} = \mathbf{0}$. So if $\exists \{\delta_0, \delta_1, \dots \delta_{p-1}\}$, not all zero, such that $\sum_{j=0}^{p-1} \delta_j \mathbf{X}_j = \mathbf{0}$, then the columns of $\mathbf{X}$ are linearly dependent. Otherwise, the columns of $\mathbf{X}$ are linearly independent. If the columns of $\mathbf{X}$ are linearly independent, then $\mathbf{X}$ is full rank.

Also recall that an inner product matrix, $\mathbf{X}'\mathbf{X}$, has non-negative eigenvalues.

The # of non-zero (strictly positive) eigenvalues of $\mathbf{X}'\mathbf{X}$ equals $r$, and if $r < p$ ($\mathbf{X}$ less than full rank), then $\mathbf{X}$ contains an exact collinearity.

Only predictor properties determine collinearity.

Exact collinearity rarely occurs, but too much collinearity causes a loss

of numerical, statistical, and scientific accuracy.

## Matrices Providing Information about Collinearity

Difficulties with $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ arise from computing cross-products. Four versions of the inner product matrix $\mathbf{X}'\mathbf{X}$ provide information about collinearity and accuracy.

## SSCP Matrix

The $p \times p$ *sums-of-squares and cross-products (SSCP) matrix* is given by

$$\mathbf{X}'\mathbf{X} = \left\{ \sum_{i=1}^{n} x_{ij} x_{ij'} \right\}.$$

Dividing by $n$ yields $\mathbf{X}'\mathbf{X}/n$, the average cross-products matrix.

## Scaled SSCP Matrix

Consider the $p \times p$ matrix

$$\boldsymbol{D}_s = \begin{bmatrix} \sum x_{i0}^2 & & & 0 \\ & \sum x_{i1}^2 & & \\ & & \ddots & \\ 0 & & & \sum x_{i,p-1}^2 \end{bmatrix},$$

which contains the diagonal elements of the SSCP matrix.

Define the $p \times p$ scaled SSCP matrix of average scaled (but not centered) cross-products as

$$(\mathbf{X}'\mathbf{X})_s = \boldsymbol{D}_s^{-0.5}(\mathbf{X}'\mathbf{X})\boldsymbol{D}_s^{-0.5}$$

This matrix has 1's on the diagonal. Examining the eigenvalues and eigenvectors of the scaled SSCP matrix is a good way to detect collinearity with the intercept.

## Covariance Matrix

Partition the covariate matrix $\mathbf{X}$ into $\begin{bmatrix} \mathbf{J}_n & \mathbf{X}^* \end{bmatrix}$, where $\mathbf{X}^*$ is the $n \times (p-1)$ matrix excluding the intercept.

The $(p-1) \times (p-1)$ covariance matrix adjusts for predictor means.

Recall that the column means of $\mathbf{X}^*$ are defined by

$$\overline{\mathbf{X}^*} = \mathbf{X}^{*'}\mathbf{J}_n/n = \left\{ \sum_{i=1}^{n} x_{ij}^*/n \right\} .$$

We define the covariance matrix $\mathbf{C}$ by

$$
\begin{aligned}
\boldsymbol{C} = \{c_{jj'}\} &= \left\{ \sum_{i=1}^{n} (x^*_{ij} - \overline{x^*}_j)(x^*_{ij'} - \overline{x^*}_{j'}) \right\} /n \\
&= \mathbf{X}^{*\prime}(\boldsymbol{I}_n - \mathbf{J}_n \mathbf{J}'_n/n)\mathbf{X}^*/n \\
&= \mathbf{X}^{*\prime}\mathbf{X}^*/n - \overline{\mathbf{X}^*}\,\overline{\mathbf{X}^*}' \\
&= \mathbf{X}^{*\prime}_c \mathbf{X}^*{}_c/n,
\end{aligned}
$$

where $\mathbf{X}^*_c = \{(x^*_{ij} - \overline{x^*}_j)\}$ contains the $p-1$ columns of centered data. Note that the diagonal of $\mathbf{C}$ contains sample variances of the predictors.

The covariance matrix contains average centered cross-products and excludes the intercept. It is also an inner product matrix as shown above.

If $\mathbf{X}$ is full rank, then $\mathbf{C}$ has rank $p-1$.

## Correlation Matrix

Extract the diagonal elements of the covariance matrix $\mathbf{C}$:

$$\boldsymbol{D}_c = \begin{bmatrix} c_{11} & & & 0 \\ & c_{22} & & \\ & & \ddots & \\ 0 & & & c_{p-1,p-1} \end{bmatrix}.$$

Define

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{D}_c^{-0.5} \boldsymbol{C} \boldsymbol{D}_c^{-0.5} \\ &= \boldsymbol{D}_c^{-0.5} \mathbf{X}_{c'}^* \mathbf{X}_c^* \boldsymbol{D}_c^{-0.5} / n \\ &= \{r_{jj'}\} = \left\{ \frac{c_{jj'}}{\sqrt{c_j c_{j'}}} \right\} = \left\{ \frac{c_{jj'}}{s_j s_{j'}} \right\}. \end{aligned}$$

Note that $c_{jj} = s_j^2$, the sample variance.

$\boldsymbol{R}$ contains average centered and scaled cross-products and is an inner

product matrix.

Also, $\mathbf{X}^*{}_c \boldsymbol{D}_c^{-0.5} = \boldsymbol{Z} = \{(x_{ij}^* - \overline{x^*}_j)/s_j\}$ equals centered and scaled data.

Examining the eigenvalues and eigenvectors of the correlation matrix is a good way to detect collinearity among predictors other than the intercept.

## Eigenanalysis!

An eigenanalysis is the best way to detect and describe collinearity.

Recall that the eigenvalues of the square matrix $\boldsymbol{A}$ are the roots the characteristic equation:

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0.$$

Equivalently, write $A\mathbf{X} = \lambda \mathbf{X}$.

Also recall that the rank of a (square) matrix equals the # of non-zero eigenvalues, so a matrix is full rank if and only if the matrix has no zero eigenvalues.

For the special case of a symmetric $\boldsymbol{A}$, we define the spectral decomposition:

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}',$$

where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ represents the diagonal matrix of eigenvalues, ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$. The $k$th column of $\boldsymbol{V}$ has an eigenvector corresponding to $k$th eigenvalue, with the eigenvectors and eigenvalues in the same order.

$\boldsymbol{V}$ is a full rank orthogonal matrix, regardless of the rank of $\boldsymbol{A}$.

Since we usually scale eigenvectors to unit length, $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$, $\boldsymbol{V}^{-1} = \boldsymbol{V}'$ and $(\boldsymbol{V}')^{-1} = \boldsymbol{V}$.

We will consider eigenanalysis of inner-product matrices like $\mathbf{X}'\mathbf{X}$, $\mathbf{C}$, and $\mathbf{R}$, which are all symmetric and positive semi-definite (so that all eigenvalues are $\geq 0$). The rank of these matrices equals the number of nonzero eigenvalues, and the number of zero eigenvalues equals the number of linear dependences. Remember that the rank of $\mathbf{X}$ is the same as the rank of $\mathbf{X}'\mathbf{X}$.

The eigenvalues of $\mathbf{X}'\mathbf{X}$, $\boldsymbol{C}$, and $\boldsymbol{R}$ are variances of linear

combinations of $X_j$'s, and eigenvectors provide regression coefficients to create new variables (variates) having the eigenvalues as variances. The relative size of eigenvalues, especially largest to smallest, indicates the amount of collinearity, while the eigenvectors allow discovery of which variables overlap.

Troubles often reduce to a scaling problem or a variable with $s^2$ near zero.

## Example: Collinearity

Consider records from one night at a hospital emergency room.

If 39 of 40 patients are male, then the indicator of gender (0 for female, 1 for male) is highly collinear with the intercept.

Consider the role of eigenvalues in determining the stability of $(\mathbf{X}'\mathbf{X})^{-1}$. Using the spectral decomposition,

$$\mathbf{X}'\mathbf{X} = \boldsymbol{V}\text{Diag}(\boldsymbol{\lambda})\boldsymbol{V}'$$

If $\mathbf{X}$ has full rank, then $(\mathbf{X}'\mathbf{X})^{-1} = \boldsymbol{V}\text{Diag}(\boldsymbol{\lambda}^{-1})\boldsymbol{V}'$. If any eigenvalues are 0, we cannot compute the usual inverse. A zero eigenvalue implies some linear combination of the predictors has a zero variance and provides no additional information.

## Principal Component Analysis

Principal component analysis describes variables using eigenanalysis.

Let $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_p \end{bmatrix}$ indicate the $p \times p$ matrix of eigenvectors of $\mathbf{X}'\mathbf{X}/n$, with each column of $\boldsymbol{V}$ a $p \times 1$ eigenvector.

Define $\mathbf{X}_* = \mathbf{X}\boldsymbol{V}$ as the $n \times p$ matrix of principal component scores. Each column represents a new variable (a variate).

The $j$th element of $k$th eigenvector ($\boldsymbol{v}_k$) provides the coefficient for the $j$th column (variable) in $\mathbf{X}$ to compute the $k$th principal component score.

Using eigenvectors of $\boldsymbol{C} = \mathbf{X}_c'\mathbf{X}_c/n$ allows computing principal component scores for centered data, while using eigenvectors of $\boldsymbol{R} = \boldsymbol{D}_c^{-0.5}\mathbf{X}_c'\mathbf{X}_c\boldsymbol{D}_c^{-0.5}/n$ allows computing principal component scores for centered and scaled data.

Except in special cases, the eigenvalues, eigenvectors, and principal component scores of $\boldsymbol{C}$, $\boldsymbol{R}$, and the SSCP matrices have no simple correspondences between them.

## Example: Eigenanalysis of Correlation Matrix for Ozone Data

```
proc princomp data=ozone;
var outdoor home time_out;
run;
*********************************************************************
                        The PRINCOMP Procedure

                      Observations          64
                      Variables              3


                         Simple Statistics

                    outdoor              home          time_out

      Mean       44.94541180       19.83328125      0.2817187500
      StD        21.90644120       11.98360958      0.2135754184



                       Correlation Matrix

         outdoor       Outdoor Ozone Concentration (ppb)
         home          Home Indoor Ozone Concentration (ppb)
         time_out      Proportion of Time Spent Outdoors
```

Correlation Matrix

| | outdoor | home | time_out |
|---|---|---|---|
| outdoor | 1.0000 | 0.5558 | 0.0782 |
| home | 0.5558 | 1.0000 | -.0071 |
| time_out | 0.0782 | -.0071 | 1.0000 |

Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 1.56035176 | 0.55838231 | 0.5201 | 0.5201 |
| 2 | 1.00196944 | 0.56429064 | 0.3340 | 0.8541 |
| 3 | 0.43767880 | | 0.1459 | 1.0000 |

```
                    Eigenvectors

outdoor       Outdoor Ozone Concentration (ppb)
home          Home Indoor Ozone Concentration (ppb)
time_out      Proportion of Time Spent Outdoors

                     Eigenvectors

                  Prin1         Prin2         Prin3

 outdoor       0.707674      0.012227      -.706434
 home          0.700809      -.139231      0.699629
 time_out      0.089803      0.990184      0.107099
```

None of the eigenvalues are too close to zero, indicating that there is not much sign of collinearity among the predictors.

Now, we also examine collinarity with the intercept by conducting an eigenanalysis of the scaled SSCP matrix.

```
data ozone; set ozone; int=1; run;

proc princomp data=ozone noint;
var int outdoor home time_out;
run;
*********************************************************************
                        The PRINCOMP Procedure

                      Observations          64
                      Variables              4


                         Simple Statistics

                  int         outdoor            home        time_out

 Mean     1.000000000     44.94541180     19.83328125    0.2817187500
 UStD     1.000000000     49.92478235     23.12405860    0.3525155138


                   Uncorrected Correlation Matrix

                      int      outdoor        home      time_out
     int           1.0000       0.9003      0.8577        0.7992
     outdoor       0.9003       1.0000      0.8966        0.7399
     home          0.8577       0.8966      1.0000        0.6832
```

| | | | | |
|---|---|---|---|---|
| time_out | 0.7992 | 0.7399 | 0.6832 | 1.0000 |

Eigenvalues of the Uncorrected Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 3.44397079 | 3.09272276 | 0.8610 | 0.8610 |
| 2 | 0.35124803 | 0.23327663 | 0.0878 | 0.9488 |
| 3 | 0.11797140 | 0.03116161 | 0.0295 | 0.9783 |
| 4 | 0.08680979 | | 0.0217 | 1.0000 |

Eigenvectors

| | Prin1 | Prin2 | Prin3 | Prin4 |
|---|---|---|---|---|
| int | 0.517493 | -.013105 | -.672666 | -.528724 |
| outdoor | 0.515072 | -.282728 | -.212037 | 0.780901 |
| home | 0.500584 | -.471532 | 0.649407 | -.324566 |
| time_out | 0.465099 | 0.835195 | 0.284308 | 0.072810 |

**Condition Number and Condition Index**

The *condition index* for the $k$th eigenvalue equals $\sqrt{\lambda_1/\lambda_k}$. The maximum condition index, called the *condition number*, $\max_k \sqrt{\lambda_1/\lambda_k}$, involves the first and last eigenvalue.

(Some authors exchange the names of condition number and condition index, so be careful!)

Note that $1 \leq \mathsf{CI}_k < \infty$, where large values of the condition index indicate greater collinearity.

**Example: Eigenanalysis of Average SSCP, Scaled SSCP, $\mathbf{C}$, and $\mathbf{R}$**
Using PROC PRINCOMP, we conduct an eigenanalysis for the ozone data. Note the VARDEF option, which is used to specify that $n$ and not $n-1$ is used in calculations for average SSCP and covariance matrices.

```
/* SSCP */
proc princomp data=ozone noint cov vardef=n;
var int outdoor home time_out;
run;
/* Scaled SSCP */
proc princomp data=ozone noint;
var int outdoor home time_out;
```

```
run;
/* Covariance Matrix */
proc princomp data=ozone noint cov vardef=n;
var outdoor home time_out;
run;
/* Correlation Matrix */
proc princomp data=ozone;
var outdoor home time_out;
run;
```

The condition indices (calculated by hand) are provided in a table.

| Matrix Type | Eigenvalue | Condition Index |
|---|---|---|
| Average SSCP | 2939.11 | 1.00 |
| | 88.99 | 5.75 |
| | 0.19 | 23.90 |
| | 0.04 | 267.81 |
| Scaled SSCP | 3.44 | 1.00 |
| | 0.35 | 3.13 |
| | 0.12 | 5.40 |
| | 0.09 | 6.30 |

| Matrix Type | Eigenvalue | Condition Index |
|---|---|---|
| Covariance | 2938.29 | 1.00 |
| | 88.99 | 5.75 |
| | 0.06 | 229.10 |
| Correlation | 1.56 | 1.00 |
| | 1.00 | 1.25 |
| | 0.44 | 1.89 |

We next consider interpretation of these values.

## Interpretation of Condition Number

One rule of thumb is that CN$> 30$ implies moderate to severe collinearity. This corresponds to a ratio of variances of roughly 1000, so single computations may involve numbers varying by 3 decimal places. Loss of precision occurs as such errors accumulate.

Note that $\lambda_k$ very near zero indicates a redundancy in the predictors, and the eigenvector can identify the culprits. The relative sizes of $\boldsymbol{v}_k = \{v_{jk}\}_{p\times 1}$ give the relative importance of the columns of $\mathbf{X}$ in determining the undesirable $k$th variate. The $k$th variate has zero variance, and hence no predictive value. We interpret the $\{v_{jk}\}$ as regression coefficients.

Consider the SSCP and scaled SSCP matrices. A $\lambda_k$ near zero indicates either collinearity with the intercept or among other predictors. Consider carefully the value of the eigenvector (with a bad condition number) weighting the intercept. A relatively large value

indicates a collinearity with the intercept.

Many simple mistakes may create collinearity with the intercept:

- using birth year rather than age,
- analyzing FVC in mL, rather than FVC in L,
- including variables with near zero variance (39/40 males),
- including redundant codings (including male and female codes), and
- some combination of the above.

$\boldsymbol{C}$ and $\boldsymbol{R}$ diagnose collinearity in predictors other than the intercept. $\boldsymbol{C}$ involves the relative scales of the variables, and $\boldsymbol{R}$ does not. Since few tests vary due to a change in scale or location, we prefer $\boldsymbol{R}$, assuming no loss of precision.

## $R_j^2$, Tolerance, and VIF

### Correlation and Collinearity

The simplest type of collinearity for variables other than the intercept is a correlation of 1.0 between two predictors.

### Example: Exact Collinearity

Let $X_1$ represent temperature in F, and let $X_2$ represent temperature in C. Then $x_{i1} = 32 + 1.8x_{i2}$.

Fitting the model $x_{i1} = \beta_0 + \beta_1 x_{i2} + \varepsilon_i$ yields $\widehat{\sigma}^2 = 0$ and $r^2(x_1, x_2) = 1$.

Suppose we use $X_j$ as the response in a regression with the $p-2$ predictors (excluding the intercept),

$$\{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\}.$$

Then we compute the *squared multiple correlation*

$$R_j^2 = R^2(X_j, \{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\}) \,.$$

Clearly $R_1^2 = R_2^2 = 1$.

Large values (close to 1) of $R_j^2$ imply worse collinearity by indicating the extent of redundancy of a predictor with the remaining ones, and values close to zero indicate little or no collinearity.

Define *tolerance* as $1 - R_j^2$. Tolerance close to 1 is good, while tolerance close to 0 is bad.

Define the *variance inflation factor* as

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}}.$$

The name comes from the fact that $\text{var}(\widehat{\beta}_j)$ is proportional to $\text{VIF}_j$.

The variance inflation factor shows how multicollinearity has increased

the instability of the coefficient estimates. The variance of the parameter estimate is larger (by the VIF) than it would be if there were no multicollinearity.

A VIF close to 1 is good, and VIF $\longrightarrow \infty$ is bad.

*Summary:*

| Good | | Diagnostic | | Bad |
|---|---|---|---|---|
| 0 | $\leq$ | $R_j^2$ | $\leq$ | 1 |
| 1 | $\geq$ | $1 - R_j^2$ | $\geq$ | 0 |
| 1 | $\leq$ | $\frac{1}{1-R_j^2}$ | $\leq$ | $\infty$ |

Eliminating a redundant variable does not reduce prediction!

As a rule of thumb,

$R_j^2 > .90$ merits some attention, and

$R_j^2 > .98$ suggests the need for removal of a useless variable.

## Example: Diagnosing Collinearity in Ozone Data

Suppose that we fit the model $y_i = \beta_0 + \beta_1 OUTDOOR_i + \beta_2 HOME_i + \beta_3 {HOME_i}^2 + \beta_4 {HOME_i}^3 + \beta_5 TIMEOUT_i + \varepsilon_i$ to the ozone data so that we can investigate whether a polynomial in home provides a better fit to the data. If so, we obtain the following regression model output.

```
data ozone; set ozone;
home2=home*home;
home3=home*home2;
run;

proc reg;
model personal=outdoor home home2 home3 time_out/tol vif;
run;
*************************************************************************

                          The REG Procedure
                            Model: MODEL1
                     Dependent Variable: personal
```

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 5109.91477 | 1021.98295 | 5.88 | 0.0002 |
| Error | 58 | 10073 | 173.67557 | | |
| Corrected Total | 63 | 15183 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 13.17860 | R-Square | 0.3366 |
| Dependent Mean | 23.54578 | Adj R-Sq | 0.2794 |
| Coeff Var | 55.97012 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 2.34851 | 8.45710 | 0.28 | 0.7822 |
| outdoor | 1 | 0.10435 | 0.09388 | 1.11 | 0.2709 |
| home | 1 | 0.56187 | 1.46423 | 0.38 | 0.7026 |
| home2 | 1 | 0.01058 | 0.07103 | 0.15 | 0.8821 |
| home3 | 1 | -0.00024100 | 0.00098237 | -0.25 | 0.8071 |
| time_out | 1 | 13.70723 | 7.95522 | 1.72 | 0.0902 |

Parameter Estimates

| Variable | DF | Tolerance | Variance Inflation |
|---|---|---|---|
| Intercept | 1 | . | 0 |
| outdoor | 1 | 0.65176 | 1.53431 |
| home | 1 | 0.00895 | 111.68575 |
| home2 | 1 | 0.00160 | 624.92088 |
| home3 | 1 | 0.00415 | 241.25289 |
| time_out | 1 | 0.95497 | 1.04715 |

Is there evidence of collinearity?

To investigate further, we conduct eigenanalyses of the scaled SSCP and correlation matrices.

```
/* Scaled SSCP matrix */
proc princomp data=ozone noint;
var int outdoor home home2 home3 time_out;
run;

/* correlation matrix */
proc princomp data=ozone;
```

```
var outdoor home home2 home3 time_out;
run;
*********************************************************************
                      The PRINCOMP Procedure

                      Observations          64
                      Variables              6


                         Simple Statistics

                        int           outdoor              home

      Mean       1.000000000       44.94541180       19.83328125
      UStD       1.000000000       49.92478235       23.12405860

                         Simple Statistics

                      home2             home3          time_out

      Mean       534.7220859       17251.08446      0.2817187500
      UStD       788.6788104       31240.89600      0.3525155138


                  Uncorrected Correlation Matrix
```

| | int | outdoor | home | home2 | home3 | time_out |
|---|---|---|---|---|---|---|
| int | 1.0000 | 0.9003 | 0.8577 | 0.6780 | 0.5522 | 0.7992 |
| outdoor | 0.9003 | 1.0000 | 0.8966 | 0.7937 | 0.7039 | 0.7399 |
| home | 0.8577 | 0.8966 | 1.0000 | 0.9459 | 0.8610 | 0.6832 |
| home2 | 0.6780 | 0.7937 | 0.9459 | 1.0000 | 0.9771 | 0.5338 |
| home3 | 0.5522 | 0.7039 | 0.8610 | 0.9771 | 1.0000 | 0.4355 |
| time_out | 0.7992 | 0.7399 | 0.6832 | 0.5338 | 0.4355 | 1.0000 |

Eigenvalues of the Uncorrected Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 4.81211368 | 3.97703127 | 0.8020 | 0.8020 |
| 2 | 0.83508241 | 0.60164583 | 0.1392 | 0.9412 |
| 3 | 0.23343658 | 0.14484901 | 0.0389 | 0.9801 |
| 4 | 0.08858757 | 0.05833043 | 0.0148 | 0.9949 |
| 5 | 0.03025714 | 0.02973452 | 0.0050 | 0.9999 |
| 6 | 0.00052262 | | 0.0001 | 1.0000 |

The PRINCOMP Procedure

Eigenvectors

| | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 | Prin6 |
|---|---|---|---|---|---|---|
| int | 0.406570 | 0.394870 | -.446481 | -.481845 | 0.489442 | -.087778 |
| outdoor | 0.428643 | 0.164398 | -.405351 | 0.785773 | -.086223 | -.007426 |
| home | 0.447257 | -.107067 | -.121952 | -.352657 | -.660918 | 0.460919 |
| home2 | 0.421044 | -.411815 | 0.117787 | -.107096 | -.154855 | -.777050 |
| home3 | 0.387170 | -.549914 | 0.264047 | 0.103256 | 0.539792 | 0.419447 |
| time_out | 0.351780 | 0.577573 | 0.733402 | 0.062342 | -.029063 | -.007115 |

The PRINCOMP Procedure

| | |
|---|---|
| Observations | 64 |
| Variables | 5 |

Simple Statistics

| | outdoor | home | home2 |
|---|---|---|---|
| Mean | 44.94541180 | 19.83328125 | 534.7220859 |
| StD | 21.90644120 | 11.98360958 | 584.3126423 |

Simple Statistics

| | home3 | time_out |
|---|---|---|

| | | |
|---|---|---|
| Mean | 17251.08446 | 0.2817187500 |
| StD | 26251.89177 | 0.2135754184 |

Correlation Matrix

| | outdoor | home | home2 | home3 | time_out |
|---|---|---|---|---|---|
| outdoor | 1.0000 | 0.5558 | 0.5730 | 0.5697 | 0.0782 |
| home | 0.5558 | 1.0000 | 0.9642 | 0.9037 | -.0071 |
| home2 | 0.5730 | 0.9642 | 1.0000 | 0.9836 | -.0181 |
| home3 | 0.5697 | 0.9037 | 0.9836 | 1.0000 | -.0116 |
| time_out | 0.0782 | -.0071 | -.0181 | -.0116 | 1.0000 |

Eigenvalues of the Correlation Matrix

| | Eigenvalue | Difference | Proportion | Cumulative |
|---|---|---|---|---|
| 1 | 3.31662552 | 2.30161843 | 0.6633 | 0.6633 |
| 2 | 1.01500708 | 0.44509651 | 0.2030 | 0.8663 |
| 3 | 0.56991057 | 0.47248772 | 0.1140 | 0.9803 |
| 4 | 0.09742285 | 0.09638887 | 0.0195 | 0.9998 |

| | | | |
|---|---|---|---|
| 5 | 0.00103398 | 0.0002 | 1.0000 |

Eigenvectors

| | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 |
|---|---|---|---|---|---|
| outdoor | 0.389979 | 0.167238 | 0.905409 | 0.012712 | -.004641 |
| home | 0.524979 | -.038764 | -.227787 | 0.749446 | 0.330663 |
| home2 | 0.539900 | -.050271 | -.226048 | -.094914 | -.803663 |
| home3 | 0.529906 | -.042101 | -.208732 | -.655080 | 0.494700 |
| time_out | 0.004661 | 0.982970 | -.183526 | -.005519 | -.006083 |

## Detecting Numerical Inaccuracy

### Which Matrix Should Be Examined?

Consider the ordered list of matrices: $\boldsymbol{R}$, $\boldsymbol{C}$, SSCP, and $\mathbf{X}$. The order corresponds to the following rank order of:

- most to least processed,
- least to most information about scaling problems,
- least to most information about location problems, and
- easiest to hardest in which to see collinearity (beyond intercept).

A general strategy for investigating collinearity is provided below.

1. Look at descriptive statistics (mean, variance) for all predictors. If necessary, change location or scale and remove variables with zero variance.

2. Examine eigenvalues and eigenvectors of scaled SSCP matrix to check for collinearity with the intercept and eliminate problems (by changing location or scale or deleting variables) if necessary.

3. Examine eigenvalues and eigenvectors of $\mathbf{R}$ to discover collinearity among other predictors.

## Location Problems

Consider a full rank GLM that includes an intercept.

In theory, slopes are location invariant. In addition, a full rank GLM which spans an intercept (and a GLH test which does not) has an $F$ statistic and p-value invariant to location. In practice, finite precision arithmetic may create violations.

Any variation in $F$ and p-value due to location implies a problem (for a model spanning an intercept and a GLH test which does not).

Only the scaled SSCP, SSCP, and $\mathbf{X}$ directly detect location problems. Eigenanalysis of the scaled SSCP matrix, means, histograms, and extreme values for $\boldsymbol{y}$ and $\mathbf{X}$ help detect location problems.

Extreme ratios of data values and SSCP or scaled SSCP eigenvalues (or elements) warn of location-induced inaccuracy.

## Scaling Problems

Slope estimates may vary due to changes in scale.

In theory, a full rank GLM which spans an intercept (and a GLH test which does not involve the intercept) has $F$ statistic and p-value invariant to change in scale of $\boldsymbol{y}$ and/or $\mathbf{X}$.

Any variation in $F$ statistic and p-value due to change of scale indicates a problem (if model spans an intercept and test does not).

$\boldsymbol{C}$, scaled SSCP, SSCP and $\mathbf{X}$ yield information about scale problems.

Extreme ratios of values of $\boldsymbol{C}$ (especially the variances) and histograms and extreme values of $\boldsymbol{y}$ and $\mathbf{X}$ can detect scale problems.

## Collinearity Problems

In theory a GLH test that does not span an intercept in a full rank GLM which spans an intercept has an $F$ statistic and p-value invariant to full rank linear transformation of $\boldsymbol{y}$ and/or the columns of $\mathbf{X}$

Any variation in $F$ statistic and p-value due to a full rank linear transformation diagnoses a problem (if the model spans an intercept and the test does not).

$\boldsymbol{R}$ demonstrates invariance to location and scale changes.

$\boldsymbol{R}$, its eigenvalues, and $R_j^2$'s detect collinearity problems.

The eigenvector for the smallest eigenvalue of $\mathbf{R}$ identifies the culprits.

## Recommendations

1. Conduct data validation as part of data entry and file creation.
2. Treat non-independence of observations, if necessary. (For example, use multivariate methods, such as repeated measures ANOVA or random effects models.)
3. Minimize location and scale problems by roughly aligning variable ranges of variables. Centering and scaling may be necessary.
4. Minimize collinearity.
   - Eliminate predictors with near zero variance.
   - Eliminate redundant variables defined by unimportance or $R_j^2$.
   - Avoid unnecessary collinearity: center polynomials or use orthogonal polynomials (Chapter 9), and use cell-mean or effect coding (Chapter 12).
5. Treat "linearity" (model specification error). Find useful predictors and transformations (Chapter 10).

6. Treat non-normality and heterogeneity. Consider transformations. See Chapter 11 for a general treatment of exploratory regression.

We will address all of these issues in the coming weeks.

To summarize, first, clean the data. Then scale, center, and code it sensibly. If necessary, transform and delete variables. Do not be satisfied with the model unless all diagnostics create no substantial cause for concern. Chapter 11 centers on using this approach while creating statistically defensible estimates and hypothesis tests.

## Next: Polynomial Regression

*Reading Assignment:*

- Muller and Fetterman, Chapter 9: "Polynomial Regression" (Required)

# Lecture 12: Polynomials and Splines

*Reading Assignment:*

- Muller and Fetterman, Chapter 9: "Polynomial Regression"

## Motivation

Response-predictor relationships are often not linear. In most cases, we want to use tests and models with maximum power to detect association between a predictor and a response. The tests and models that we have discussed so far assume a linear relationship between $E(\mathbf{y})$ and $\mathbf{x}$, and power for detecting an association between $\mathbf{y}$ and $\mathbf{x}$ will be reduced when this is not true. (A linear model may have good power if the trend is nonlinear but still monotone, but power will likely be terrible for U-shaped or umbrella-shaped relationships.)

A simple way to capture non-linear relationships between an exposure and response of interest is to add polynomial terms to a linear model. With $d$ distinct $(x, y)$ pairs, a polynomial of order at most $d - 1$ will pass exactly through all the points. Consider the following example, in which $\mathbf{x} = (1, 2, 3, 4, 5)'$ and $\mathbf{y} = (5, 0, 2, 1, 4)'$.

Why wouldn't we always want to use this type of model, which perfectly fits the data?

## Polynomial Models

| | Order | Regression Model | Shape |
|---|---|---|---|
| 0th | Zero | $y_i = \beta_0 + \varepsilon_i$ | Horizontal Line |
| 1st | Linear | $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ | Line |
| 2nd | Quadratic | $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$ | Parabola |
| 3rd | Cubic | $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$ | Ogive etc. |

The number of points of inflection is one fewer than the number of monotone pieces of the curve, which is the order (degree) of the polynomial. The maximum order that can be fitted is one less than the number of distinct $X$ values, but one should be able to defend the use of any model with higher than a 3rd order polynomial due to difficulties in interpretation.

We will use the GLM for estimation of parameters, but how?

## Example: Romanesque Cathedrals in England

Consider the following data collected by Stephen J. Gould on the dimensions of medieval English cathedrals. (The Romanesque period in English architecture roughly dates to 1066-1180.)

| Cathedral | Nave Height (ft) | Total Length (ft) |
|---|---|---|
| Durham | 75 | 502 |
| Canterbury | 80 | 522 |
| Gloucester | 68 | 425 |
| Hereford | 64 | 344 |
| Norwich | 83 | 407 |
| Peterborough | 80 | 451 |
| St. Albans | 70 | 551 |
| Winchester | 76 | 530 |
| Ely | 74 | 547 |

The nave height is the height of the center of the church, as seen (for Canterbury Cathedral) on the following page.

We can fit a polynomial model, say a third order polynomial, to the cathedral data, treating length as the dependent variable and height as the independent variable. The $\mathbf{X}$ matrix is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 75 & 75^2 & 75^3 \\ 1 & 80 & 80^2 & 80^3 \\ 1 & 68 & 68^2 & 68^3 \\ 1 & 64 & 64^2 & 64^3 \\ 1 & 83 & 83^2 & 83^3 \\ 1 & 80 & 80^2 & 80^3 \\ 1 & 70 & 70^2 & 70^3 \\ 1 & 76 & 76^2 & 76^3 \\ 1 & 74 & 74^2 & 74^3 \end{bmatrix}.$$

Variable added-last tests for each term are not recommended for

model selection. (Why?)

To test association (i.e., is $\mathbf{x}$ related to $\mathbf{y}$?) when linearity is not assumed, we will conduct a *groupwise* test of all terms involving the predictor.

We fit a cubic model to the cathedral data below.

```
data romanesque; set romanesque;
heightsq=height*height; heightcu=heightsq*height;
run;
proc reg data=romanesque;
model length=height heightsq heightcu/tol vif ss1;
output out=out pred=pred;
run;
*********************************************************************
                          The REG Procedure
                            Model: MODEL1
                     Dependent Variable: length

                         Analysis of Variance

                                  Sum of          Mean
 Source                  DF      Squares        Square  F Value  Pr > F

 Model                    3        34106         11369     7.41  0.0275
 Error                    5   7676.04705    1535.20941
 Corrected Total          8        41782


         Root MSE              39.18175    R-Square     0.8163
         Dependent Mean       475.44444    Adj R-Sq     0.7061
```

```
        Coeff Var              8.24108
```

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -14512 | 30517 | -0.48 | 0.6544 |
| height | 1 | 478.38312 | 1253.74118 | 0.38 | 0.7185 |
| heightsq | 1 | -4.69654 | 17.10590 | -0.27 | 0.7946 |
| heightcu | 1 | 0.01324 | 0.07752 | 0.17 | 0.8711 |

Parameter Estimates

| Variable | DF | Type I SS | Tolerance | Variance Inflation |
|---|---|---|---|---|
| Intercept | 1 | 2034427 | . | 0 |
| height | 1 | 3035.20888 | 0.00000317 | 315582 |
| heightsq | 1 | 31026 | 7.844759E-7 | 1274736 |
| heightcu | 1 | 44.79835 | 0.00000310 | 323080 |

We consider three basic inferential questions.

1. Are height and length related?

2. Is a cubic model significantly better than the quadratic model? If not, is a quadratic model significantly better than a linear model?

3. Should we have considered even higher-order polynomial terms?

How do you interpret the tolerance and VIF values?

Tolerance is a measure of collinearity, which equals to 1-$R^2$ of predicting this covariate with all the other covariates. VIF $=$ 1/tolerance.

The following code may be used to obtain a plot of observed and predicted length values for the cubic model.

```
symbol1 color=black i=none v=star;
symbol2 color=black v=point line=1 i=RC;
/* RL for linear, RQ for quadratic */
proc gplot data=out;
plot length*height=1 pred*height=2/overlay;
run;
```

What do you think about the model fit based on the observed and predicted nave lengths? What do you think about inferences outside the range of the data?

## Limitations of Natural Polynomials

Fitting higher-order natural polynomial models can be dangerous! Because independent variables in a polynomial model are functions of the same basic variable $x$, they are correlated, and computational difficulties due to collinearity may result. In fact, collinearity is almost certainly present if the order of the polynomial is high. Techniques like centering and the use of orthogonal polynomials will help, though there is no remedy for difficulty of interpretation for high-order polynomials in most settings.

With a polynomial model, extrapolation beyond the range of the data is even more dangerous than usual.

## Orthogonal Polynomials

*Orthogonal polynomials* provide one good solution to numerical problems in a polynomial regression by using an alternate coding scheme. Natural polynomials yield the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i,$$

while orthogonal polynomials yield the model

$$y_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2} + \gamma_3 z_{i3} + \delta_i.$$

The orthogonal polynomial scores $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$ are the solution to a system of equations under the constraints that the new predictors $(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$

1. contain the same information as the original variables $(\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3)$,
2. are linear combinations of original data,
3. are mean zero (except for the intercept),

4. are mutually orthogonal (uncorrelated), and
5. $\mathbf{z}_1$ captures all information in $\mathbf{x}$ beyond location of response (intercept), $\mathbf{z}_2$ captures all information in $\mathbf{x}^2$, beyond linear term in $\mathbf{x}$ and the intercept, etc.

Finding the values of $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ corresponds to solving a system of simultaneous equations for each subject:

$$
\begin{aligned}
z_{i1} &= a_{01} \;+\; a_{11}x_i \\
z_{i2} &= a_{02} \;+\; a_{12}x_i \;+\; a_{22}x_i^2 \\
&\vdots \\
z_{ip-1} &= a_{0p-1} \;+ \sum_{j=1}^{p-1} a_{jp-1}(x_i)^j \,.
\end{aligned}
$$

Clearly, the new variables are linear combinations of the original ones. We may write

$$
\begin{aligned}
x_i &= b_{01} \;+\; b_{11} z_{i1} \\
x_i^2 &= b_{02} \;+\; b_{12} z_{i1} \;+\; b_{22} z_{i2} \\
&\vdots \\
x^{p-1} &= b_{0p-1} \;+ \sum_{j=1}^{p-1} b_{jp-1} z_{ij}.
\end{aligned}
$$

Thus we may (without losing any information) write either

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_{p-1} x_i^{p-1} + \varepsilon_i$$

or

$$y_i = \alpha_0 + \alpha_1 z_{i1} + \ldots + \alpha_{p-1} z_{ip-1} + \varepsilon_i.$$

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ will not be numerically identical, but the overall $F$ tests from these models are exactly the same. In addition, testing $H_0 : \alpha_k = 0$ for the orthogonal polynomial model is equivalent to an added-in-order test of $H_0 : \beta_k = 0$ in the natural polynomial

model.

If the predictor values are equally spaced and there are an equal number of observations at each of $d$ values, finding the orthogonal predictors reduces to finding $d-1$ vectors, each of length $d$. The $d-1$ vectors are known as the orthogonal polynomial coefficients. The table in Appendix B.10 of Muller and Fetterman may be used to obtain coefficients if predictors are equally spaced and each predictor value has the same number of replicates.

## Other Ways to Improve Fit of Natural Polynomial Models

- Scaling: For example, if a weight variable is measured in grams with a range of 1000-2000, use kg with a range of 1-2. (Presence of the intercept leads to preferring a range of 1-10 to avoid collinearity.)
- Centering: Often dramatically reduces collinearity (always eliminates any with intercept) and is strongly recommended as habitual technique.
- Pseudo-Centering: Some "nice" number often provides most of the numerical advantage of centering, while simplifying interpretation. So if a predictor is weight or blood pressure, we might pseudo-center it at a "healthy" value of the predictor (average in a healthy population say).

## Model Selection in Polynomial Regression

Muller and Fetterman prefer *backwards selection*, starting at the largest desirable polynomial model (they recommend starting at a cubic) and evaluating smaller ones with added-in-order tests. They recommend stopping at the highest order polynomial that is significant and including all lower order terms. Model diagnostics are very important in polynomial models as in all other regression models.

## More Flexible Models

Often, $y$ does not behave linearly in all the predictors. The simplest way to describe a nonlinear effect of $x$ is to include polynomial terms in the model. However, nonlinear effects may not follow polynomial forms. In such cases, transformations might be able to induce linearity, but often the transformation is not known or does not exist.

### Linear Splines

*Spline functions* are piecewise polynomials used to fit curves. Within intervals of $x$, they are polynomials, and they are connected across the different intervals. The most simple type of spline function is a *linear spline function*, which is a piecewise linear function. Linear spline functions may also be fit in the framework of the GLM.

Divide the $x$ axis into intervals with endpoints $k_1, k_2, \ldots, k_K$, called *knots*. The linear spline function is given by

$$\begin{aligned} E(\mathbf{y}) &= \beta_0 + \beta_1 \mathbf{x} + \beta_2(\mathbf{x} - k_1 \mathbf{J}_n)_+ + \beta_3(\mathbf{x} - k_2 \mathbf{J}_n)_+ \\ &\quad + \ldots + \beta_{K+1}(\mathbf{x} - k_K \mathbf{J}_n)_+, \end{aligned}$$

where

$$(u)_+ = \begin{cases} u & u > 0 \\ 0 & u \leq 0. \end{cases}$$

The number of knots varies depending on the amount of data available.

We can rewrite the linear spline function as follows:

$$E(y_i) = \begin{cases} \beta_0 + \beta_1 x_i & x_i \leq k_1 \\ \beta_0 + \beta_1 x_i + \beta_2(x_i - k_1) & k_1 < x_i \leq k_2 \\ \beta_0 + \beta_1 x_i + \beta_2(x_i - k_1) + \beta_3(x_i - k_2) & k_2 < x_i \leq k_3 \\ \vdots & \\ \beta_0 + \beta_1 x_i + \beta_2(x_i - k_1) + \ldots + \beta_{K+1}(x_i - k_K) & k_K < x_i. \end{cases}$$

We can fit this model in SAS by creating new variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K)$ where

$$\begin{aligned} \mathbf{x}_1 &= \mathbf{x} \\ \mathbf{x}_2 &= (\mathbf{x} - k_1)_+ \\ \vdots & \quad \vdots \\ \mathbf{x}_K &= (\mathbf{x} - k_K)_+. \end{aligned}$$

Then, we can fit the model

$$E(\mathbf{y}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots + \beta_{K+1} x_K$$

in any standard software package.

To test linearity in $x$, we simply test the hypothesis

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_{K+1} = 0.$$

To construct a linear spline with knots at $k_1 = 70$ and $k_2 = 77$ for the cathedral data, we use the following $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & 75 & 5 & 0 \\ 1 & 80 & 10 & 3 \\ 1 & 68 & 0 & 0 \\ 1 & 64 & 0 & 0 \\ 1 & 83 & 13 & 6 \\ 1 & 80 & 10 & 3 \\ 1 & 70 & 0 & 0 \\ 1 & 76 & 6 & 0 \\ 1 & 74 & 4 & 0 \end{bmatrix}.$$

Then we can fit the model using the following SAS code.

```
data romanesque; set romanesque;
ht_a=0;
ht_b=0;
if height>70 then ht_a=height-70;
if height>77 then ht_b=height-77;
run;

proc reg;
model length=height ht_a ht_b;
run;
*************************************************************************
```

The REG Procedure
Model: MODEL1
Dependent Variable: length

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 35677 | 11892 | 9.74 | 0.0157 |
| Error | 5 | 6105.05660 | 1221.01132 | | |
| Corrected Total | 8 | 41782 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 34.94297 | R-Square | 0.8539 |
| Dependent Mean | 475.44444 | Adj R-Sq | 0.7662 |
| Coeff Var | 7.34954 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | -1738.74528 | 534.64321 | -3.25 | 0.0226 |
| height | 1 | 32.35377 | 7.92328 | 4.08 | 0.0095 |

| | | | | | |
|---|---|---|---|---|---|
| ht_a | 1 | -31.48585 | 12.78943 | -2.46 | 0.0571 |
| ht_b | 1 | -19.83333 | 12.62576 | -1.57 | 0.1770 |

Parameter Estimates

| Variable | DF | Type I SS |
|---|---|---|
| Intercept | 1 | 2034427 |
| height | 1 | 3035.20888 |
| ht_a | 1 | 29629 |
| ht_b | 1 | 3012.97872 |

The plot of fitted values vs. height is below.

Test linearity using this model. How do you interpret the parameter estimates?

## Cubic Spline Functions

Although linear splines are simple and are good approximations for some relationships, they are not smooth, and they will not fit curved functions very well. To do so, we must use piecewise polynomials of higher orders. Cubic polynomials have nice properties and do a good job of fitting sharp curves. To make cubic splines smooth at the knots, we force the first and second derivatives of the function to agree at the knots.

## Spline Functions in SAS

PROC GAM (generalized additive model) in SAS allows us to fit spline functions for covariates. These splines are slightly different from cubic spline functions but are based on the same ideas that we have discussed.

## Categorization of Predictors

In epidemiology, categorizing continuous predictors is almost a default practice. The thought is that creating categories of exposure (or dichotomizing exposure into "exposed" and "unexposed" groups) avoids assumptions of linearity. Although categorization does make interpretation simple, it can make unnatural assumptions about the exposure effect and may also lead to power losses.

For example, suppose that we decide to divide churches into three groups: short ($\leq 70$ feet), medium ($> 70$ and $\leq 77$ feet), and tall ($> 77$ feet). We can then define two variables to indicate whether churches are medium or tall:

$$
\begin{aligned}
medium_i &= \begin{cases} 1 & 70 < height \leq 77 \\ 0 & \text{otherwise} \end{cases} \\
tall_i &= \begin{cases} 1 & height > 77 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

We can then fit the model

$$E(length) = \beta_0 + \beta_1 medium_i + \beta_2 tall_i + \varepsilon_i,$$

$i = 1, \ldots, n$. In this model, we have

$$
\begin{aligned}
E(length|\text{short nave}) &= \beta_0 \\
E(length|\text{medium nave}) &= \beta_0 + \beta_1 \\
E(length|\text{tall nave}) &= \beta_0 + \beta_2.
\end{aligned}
$$

What does this model assume about the effect of increasing nave height among churches with short naves?

Below are SAS code and output from fitting this model.

```
data romanesque; set romanesque;
medium=1; tall=0;
if height<70.1 then medium=0;
if height>77 then medium=0;
if height>77 then tall=1;
run;

proc reg;
model length=medium tall/ss1;
run;
*********************************************************************
                          The REG Procedure
                            Model: MODEL1
                      Dependent Variable: length

                         Analysis of Variance

                                  Sum of          Mean
 Source                  DF      Squares        Square  F Value  Pr > F

 Model                    2        12254   6126.77778     1.24  0.3530
 Error                    6        29529   4921.44444
 Corrected Total          8        41782
```

| | | | |
|---|---|---|---|
| Root MSE | 70.15301 | R-Square | 0.2933 |
| Dependent Mean | 475.44444 | Adj R-Sq | 0.0577 |
| Coeff Var | 14.75525 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 440.00000 | 40.50286 | 10.86 | <.0001 |
| medium | 1 | 86.33333 | 57.27969 | 1.51 | 0.1825 |
| tall | 1 | 20.00000 | 57.27969 | 0.35 | 0.7389 |

Parameter Estimates

| Variable | DF | Type I SS |
|---|---|---|
| Intercept | 1 | 2034427 |
| medium | 1 | 11654 |
| tall | 1 | 600.00000 |

A plot of the model fit is provided.

Fit of model E(y)=b0+b1*medium+b2*tall
cath$length
350
400
450
500
550
65
70
75
80
cath2$ht

Test association in this model. In addition, give proper interpretations of all parameter estimates, and provide estimated conditional means for cathedrals with nave heights of 69.9 and 70.1 feet, respectively.

## Nonparametric Regression

*Nonparametric smoothers* are tools that help determine the shape of the relationship between variables. These tools work best when interest is in one continuous predictor and one continuous response at a time.

### Moving Average

The most simple nonparametric smoother is a *moving average*. Suppose our data are $\mathbf{x} = (1, 2, 3, 5, 8)'$ and $\mathbf{y} = (2.1, 3.8, 5.7, 11.1, 17.2)'$. To smooth their relationship, we could estimate $E(Y|X = 2)$ by $\frac{2.1+3.8+5.7}{3}$. This is problematic at the upper and lower limits $X$, and estimates are sensitive to the interval width, which here is 3 data points. Moving average smoothers are also called moving flat line smoothers.

## Loess

A moving least squares linear regression smoother is superior to the moving average approach. The smoother *loess* is the most popular smoother of this type. To get smoothed values of $E(Y)$ at $X = x$, we take all data with $X$ values within a suitable interval (often $\frac{2}{3}$ of data points) about $x$, and then we fit a linear regression using only these points. The predicted value from this regression at $X = x$ will be used as our estimate of $E(Y|X = x)$. Then, we move to the next observed value of $X$ and repeat the process.

Actually, *weighted least squares* estimates are typically used so that points closest to $X = x$ are given the most weight, and points further away are given less weight. For this reason, *loess* is called a locally weighted least squares method. Loess calculates a smoothed value for $E(Y)$ at each observed value of $X$, and estimates for other $X$'s are obtained by interpolation.

Next is a plot of the model fit using nonparametric regression with a loess smoother for nave height. Like splines, loess smoothers may be fit in PROC GAM.

## Summary

Relationships between a response and predictor are often not linear. Nonlinear effects can be accommodated in the GLM framework by using

- polynomial terms (orthogonal polynomials are more stable than natural polynomials),
- transformations (often not known in advance),
- simple splines, or
- categorization of predictors (power loss; interpretation difficulties).

Smoothing splines and loess smoothers are useful in determining shapes of relationships and may be the best choice for presenting analysis results if relationships are highly nonlinear or when thresholds are of interest. However, such smoothers are often not the most powerful approach when relationships are monotone or can be well-represented by lower-order polynomials.

## Next: Transformations

*Reading Assignment:*

- Muller and Fetterman, Chapter 10: "Transformations"
- Weisberg, Chapter 8: "Transformations"

# Lecture 13: Transformations

*Reading Assignment:*

- Muller and Fetterman, Chapter 10: "Transformations"
- Weisberg, Chapter 8: "Transformations"

Transformation of the response and/or predictor variables may correct violations of homogeneity, linearity, and Gaussian distribution of errors (often these three assumptions stand or fall as a group).
Transformation may also simplify a model by linearizing a relationship.

Using transformations often boils down to trial and error, and we use diagnostics on the residuals to gauge the value of a transformation.

Always bear in mind that nonlinear models are part of the complete picture and may be the best alternative.

For regression, we consider three groups of transformations: linear, monotone, or non-monotone. This grouping is used to indicate invariance of statistical properties. For a $T$ test, means and variances change under linear transformations, but the test statistic and p-value do not. Test statistics and p-values for many non-parametric methods do not change with monotone transformations of the data.

A transformation has statistical benefit or cost only if it changes probabilities (and hence inferences).

Transformations may help in model fitting and may provide scientific insight. If height$^2$ works better than height, then perhaps the true relationship involves surface area.

## Variance Stabilizing Transformations of the Response

*Variance stabilizing transformations* are useful in some cases for treating heteroscedasticity.

When observations are amounts or measurements (think of a ratio scale variable), the standard deviation is often proportional to the mean. As an example, think of counting the money in a coke machine versus counting all the money in a Wachovia office.

Standard variance stabilizing transformations include the following.

| Data | Distribution | Transformation |
|---|---|---|
| Count | Poisson | $\sqrt{y_i}$ |
| Amount | Gamma | $\log y_i$ |
| Proportion | Binomial/$n$ | $\sin^{-1}(\sqrt{y_i})$ |

Despite these transformations, there are much better modeling strategies for count data and proportions. *Generalized Linear Models* are more appropriate for counts and proportions/percentages since they accommodate data from distributions other than the normal distribution.

We will devote most of our energy to discussion of *linearizing* transformations.

**log transformation**

The natural logarithm is useful when one expects the effect to be proportional to the response. For example, consider a model with one predictor, $x$ in which the response is expected to increase $100\rho$ percent for each 1-unit increase in $x$. In addition, suppose that the error, $\delta$, is multiplicative. Then we can write the model

$$y = \gamma(1+\rho)^x \delta.$$

As you can see, when $x = 0$, $y = \gamma\delta$, and when $x = 1$, $y = \gamma\delta + \gamma\rho\delta$, which is a $100\rho\%$ increase!

Taking logs on both sides, we have

$$\begin{aligned} \log(y) &= \log(\gamma) + x\log(1+\rho) + \log(\delta) \\ &= \alpha + \beta x + \varepsilon, \end{aligned}$$

where $\alpha = \log(\gamma)$, $\beta = \log(1+\rho)$, and $\varepsilon = \log(\delta)$. Thus taking logs transforms the complex multiplicative model into a simple linear model.

Solving for $\rho$ in terms of $\beta$, we have $\rho = \exp(\beta) - 1$. So a one-unit increase in $x$ corresponds to a $100(\exp(\beta) - 1)\%$ increase in $y$. For small $\beta$, say $|\beta| < 0.10$, you can interpret $\widehat{\beta}$ as a $100\widehat{\beta}$ percent effect on the response.

## Power Transformation of the Response

### Box-Cox Transformations

The square root transformation involves taking $y^{\frac{1}{2}}$. We may also wish to consider other powers of $y$ in models of the form $y_i^{\pi} = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$. Although we can fit this model for various values of $\pi$, it is difficult to decide which transformation to use since the models are not directly comparable (the SSE's are not in the same units). To solve this problem, Box and Cox (1964) introduced a family of transformations of the response variable:

$$Y_i(\pi) = \begin{cases} \frac{Y_i^{\pi} - 1}{\pi \cdot Y^{*(\pi-1)}} & \pi \neq 0 \\ Y^* \cdot \ln(Y_i) & \pi = 0\,. \end{cases}$$

where $Y^* = \left(\prod_{i=1}^{N} Y_i\right)^{1/N}$, the geometric mean of the response, and the natural logarithm for $\pi = 0$ reflects the limit as $\pi \longrightarrow 0$.

This basically corresponds to the transform $y^\pi$ for $\pi \neq 0$ and $\log(y)$ for $\pi = 0$, except that the Box-Cox transformation puts the SSE's on the same scale so that they can be compared.

Box and Cox recommended choosing $\pi$ to minimize the SSE.

Below is a simple example to illustrate the Box-Cox transformation. Data were generated from the model $y = e^{x+\epsilon}$ where $\epsilon \sim N(0, 1)$. The transformed data can be fit with a linear model $\log(y) = x + \epsilon$. SAS code is below:

```
data x;
   do x = 1 to 8 by 0.025;
      y = exp(x + normal(7));
      output;
      end;
   run;

proc transreg data=x details;
   title2 Defaults;
   model boxcox(y) = identity(x);
   run;
```

```
   The TRANSREG Procedure
Transformation Information for BoxCox(y)
Lambda      R-Square    Log Like
-3.00          0.03    -4601.01
-2.75          0.04    -4266.08
-2.50          0.04    -3934.11
-2.25          0.05    -3605.75
-2.00          0.06    -3281.88
-1.75          0.07    -2963.74
-1.50          0.10    -2653.14
-1.25          0.14    -2352.72
-1.00          0.21    -2066.32
-0.75          0.34    -1799.25
-0.50          0.52    -1558.55
-0.25          0.71    -1360.28
0.00           0.79    -1275.31 <- best lambda
0.25           0.70    -1382.62
0.50           0.51    -1589.03
0.75           0.34    -1834.53
1.00           0.22    -2105.88
1.25           0.15    -2397.35
1.50           0.11    -2704.64
1.75           0.08    -3024.24
2.00           0.06    -3353.38
2.25           0.05    -3689.91
2.50           0.04    -4032.18
```

```
2.75          0.03    -4378.97
3.00          0.03    -4729.37
```

Regression assumption diagnostics must always be examined as well.

Note: all of these transformations are not well-defined if $y$ is not strictly positive. In such a case, you may encounter problems for $\pi \leq 0$.

## Ladder of Power Transformations

Alternatively, the "ladder of power transformations" below can be used to guide the choice of transformations.

Half-Steps on the Ladder of Power Transformations

| $\pi$ | Transform | | | Description |
|---|---|---|---|---|
| | $\vdots$ | | | |
| $-2$ | $y^{-2}$ | | | |
| $-3/2$ | $y^{-3/2}$ | | | |
| $-1$ | $y^{-1}$ | | | reciprocal |
| $-1/2$ | $y^{-1/2}$ | $=$ | $1/\sqrt{y}$ | |
| "0" | $\lim_{\pi \to 0} y^{\pi}$ | $=$ | $\ln y$ | |
| $1/2$ | $y^{1/2}$ | $=$ | $\sqrt{y}$ | square root |
| $1$ | $y^{1}$ | | | identity |
| $3/2$ | $y^{3/2}$ | | | |
| $2$ | $y^{2}$ | $=$ | | square |
| | $\vdots$ | | | |

A crude but effective selection strategy is outlined below.

1. Fit the proposed model and check studentized residuals, including the R/P plot, the frequency histogram for skewness, and a test of Gaussian distribution. Multi-modality suggests an important predictor (or predictors) is missing from the model. If the residuals look good, keep the current model.

2. Based on the R/P plot and frequency histogram, select a grid of $\pi$ values. Positive skewness in the histogram or a fan shape opening to the right on the R/P plot implies moving up the ladder from the value 1 (choosing values of $\pi < 1$), while negative skewness or a fan shape opening to the left on the R/P plot implies moving down ($\pi > 1$). Ratio scale variables often require $\pi < 1$.

3. Check studentized residuals for each value of $\pi$ and examine other appropriate diagnostics, including tests of the Gaussian distribution of residuals.

4. Note that if you pick $\pi < 0$, the ladder direction reverses due to

the reciprocal transformation in addition to the power transformation. So if $\pi < 0$ for the current model, an R/P plot with a fan opening to the right implies moving down the ladder, and a fan opening to the left implies moving up the ladder. For example, consider the following R/P plot for $\pi = 2$ in the ozone data.

In this case, the fan opens to the right, and we move up the ladder.

5. To select among models with residuals that look ok, choose the

model with the smallest SSE.

6. Failure to obtain an acceptable model implies a more serious problem such as inadequate predictors or an entirely inappropriate model.

## Example: Power Transformations for Ozone Data

The SAS code below can be used to program the Box-Cox transformation for the ozone data for selected values of $\pi$ in the range $[-2, 2]$.

```
data ozonex;
merge gmean ozone;
y_2=((personal**2)-1)/(2*(geomeany**(2-1)));
y_1_5=((personal**1.5)-1)/(1.5*(geomeany**(1.5-1)));
y_1=((personal**1)-1)/(1*(geomeany**(1-1)));
y_5=((personal**0.5)-1)/(0.5*(geomeany**(0.5-1)));
y_0=geomeany*log(personal);
y_m5=((personal**(-0.5))-1)/(-0.5*(geomeany**(-0.5-1)));
y_m1=((personal**(-1))-1)/(-1*(geomeany**(-1-1)));
y_m1_5=((personal**(-1.5))-1)/(-1.5*(geomeany**(-1.5-1)));
y_m2=((personal**(-2))-1)/(-2*(geomeany**(-2-1)));
run;
```

```
proc glm data=ozonex;
model y_2=outdoor home time_out;
run;

proc glm data=ozonex;
model y_1_5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_1=outdoor home time_out;
run;

proc glm data=ozonex;
model y_5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_0=outdoor home time_out;
run;


proc glm data=ozonex;
```

```
model y_m5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m1=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m1_5=outdoor home time_out;
run;

proc glm data=ozonex;
model y_m2=outdoor home time_out;
run;
```

In the interest of time, we just look at the $SSE$ for each model in the table below.

SSE for Ozone Data

| $\pi$ | $SSE$ |
|---|---|
| 2.0 | 37788 |
| 1.5 | 17671 |
| 1.0 | 10148 |
| 0.5 | 8578 |
| 0.0 | 15126 |
| -0.5 | 69083 |
| -1.0 | 637807 |
| -1.5 | 8767407 |
| -2.0 | 152261093 |

Let's examine the residuals for the untransformed data and for the square root transformed data. For the untransformed data, we have

and for the square root transformation, we have

Which transformation is best?

Alternatively, we can try the following SAS code:

```
proc transreg data=ozone ss2 details;
title2 Defaults;
model boxcox(personal) =
               identity(outdoor home time_out);
run;

Transformation Information
for BoxCox(personal)

Lambda      R-Square    Log Like
-3.00          0.05    -667.601
-2.75          0.05    -617.486
-2.50          0.05    -568.075
-2.25          0.05    -519.492
-2.00          0.05    -471.896
-1.75          0.05    -425.492
-1.50          0.05    -380.551
-1.25          0.06    -337.442
-1.00          0.07    -296.686
-0.75          0.09    -259.033
```

```
-0.50          0.14    -225.559
-0.25          0.19    -197.696
0.00          0.26    -176.954
0.25          0.31    -164.160
0.50          0.34    -158.805 <
0.75          0.34    -159.364
1.00          0.33    -164.183
1.25          0.31    -171.987
1.50          0.29    -181.931
1.75          0.27    -193.471
2.00          0.25    -206.253
2.25          0.23    -220.036
2.50          0.21    -234.650
2.75          0.19    -249.966
3.00          0.17    -265.885
```

## Comments

- Regression assumption diagnostics should guide model choice.
- Remember that Gaussian distribution, homogeneity, and linearity often stand or fall together.
- Aesthetics often dictate using parallel transformations, as for a baseline covariate (if we choose to transform personal ozone exposures, then we may also wish to transform home and outdoor exposures using the same transformation).
- Choosing some transformations (say $y^{\frac{3}{2}}$) will make models difficult to interpret.
- Ratio scale variables often need transformation.
- The nature of any zeros affects the type of transformation. Log and reciprocal transformations often work well for blood, urine, or water assays of natural compounds with non-zero background levels. Values below a detection threshold that are recorded as

zero are instead informatively censored and not missing or equal to zero. Non-natural compounds, such as some pharmaceuticals in blood, have true zeros.

- Some transformations may not work well for some data (we do not wish to take the square root of a negative number or the log of zero).

Avoid back transformation because it may lead to potential bias. For example, a scientist was studying a population with a chronic disease. A depression scale was used as the response in a GLM. Using the strategy described above, the statistician chose $\sqrt{Y}$. The scientist complimented the statistician on the analysis. However, in the draft manuscript, the scientists included means, standard deviations, and back transformed predicted values (in the original metric). The statistician convinced the scientist to stick with transformed data. Why?

Nonlinear functions usually imply

$$E[f(X)] \neq f(E[X]).$$

For example, from Jensen's inequality we know that

$$E[Y^2] > (E[Y])^2.$$

## Transforming Predictor Variables

Box and Tidwell suggest the following procedure to check whether a predictor should be transformed. To test whether a transform $x^\lambda$ should be used in the model,

1. add the covariate $a_i = x_i \log(x_i)$ to a model already containing $x_i$
2. let $\widehat{\gamma}$ be the estimated coefficient of $a_i$ and test to see if it is significantly different from zero (using a t-test say)
   (a) if not significant, no transform is needed
   (b) if significant, a preliminary guess at the proper transform is given by $\widehat{\lambda} = \frac{\widehat{\gamma}}{\widehat{\beta}} + 1$, where $\widehat{\beta}$ is the estimated coefficient of $x$ in the model containing both $x$ and $a$.

Why does this technique work?

Consider a true model,

$$y = \alpha + \beta x^{\lambda} + \varepsilon.$$

We will use Taylor's theorem to expand $x^{\lambda}$ around $\lambda = 1$. Recall from calculus that Taylor's theorem says

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \ldots.$$

In addition, recall that $\frac{\partial}{\partial x} c^x = c^x \log(c)$ for $c > 0$. Using these facts,

$$\begin{aligned} f(\lambda) &= x^{\lambda}, & \frac{\partial}{\partial \lambda} f(\lambda) &= x^{\lambda} \log(x) \\ f(\lambda = 1) &= x, & f'(\lambda = 1) &= x \log(x) \end{aligned}$$

Using a first-order Taylor series expansion with $b = \lambda$, $a = 1$,

$$x^{\lambda} \approx x + (\lambda - 1) x \log(x).$$

Substituting this into the above model, we have

$$\begin{aligned} y &= \alpha + \beta(x + \lambda x \log(x) - x \log(x)) + \varepsilon \\ &= \alpha + \beta x + \beta(\lambda - 1) x \log(x) + \varepsilon \\ &= \alpha + \beta x + \gamma x \log(x) + \varepsilon, \end{aligned}$$

where $\gamma = \beta(\lambda - 1)$. If $\gamma$ is not significant, then either $\beta = 0$ (no effect of the covariate $x$) or $\lambda = 1$ (no need for a transformation). If $\gamma$ is significant, then we solve to get $\lambda = \frac{\gamma}{\beta} + 1$.

## Next: Model Selection

*Reading Assignment:*

- Muller and Fetterman, Chapter 11: "Selecting the Best Model"
- Weisberg, Chapter 10: "Model Selection"

# Lecture 14: Selecting the Best Model

*Reading Assignment:*

- Muller and Fetterman, Chapter 12: "Selecting the Best Model" (Required)

*Selecting the best model*, while perhaps one of the most common data analysis tasks, is an *exploratory* activity that is usually accompanied by type I error rate inflation (every time we compare two models, we compromise the type I error rate of the model eventually chosen). A *confirmatory* analysis requires specifying the variables, model, and tests without knowledge of the data at hand. (This is typically the strategy used for FDA drug approval, a situation in which a great deal is known about the action of the drug through clinical trials and studies of related drugs so that such a model may be specified in advance.)

Often, investigators have many potential predictors and wish to find the best subset, and in such situations, it is important to bear in mind that p-values from hypothesis tests must be interpreted with caution in an exploratory analysis. (Alternatively, one may wish to use a correction such as the Bonferroni correction for all hypothesis tests conducted.)

## Model Selection Strategy Overview

Assuming that our goal is to use purely exploratory analysis to find the best-fitting model for the data at hand, our model selection strategy consists of four steps.

1. Specify the maximum model under consideration.
2. Specify a criterion for model selection.
3. Specify a strategy for applying the criterion.
4. Conduct the analysis.

In addition, if we wish to have confidence in our model for data outside the current sample, we should evaluate the reliability of the model chosen. In a purely exploratory analysis, one often simply interprets significance of coefficients in the final model with caution, refraining from using $p \leq 0.05$ after conducting a series of tests on the data in order to choose that final model. (The practice of doing this anyway is widespread.) However, other methods may be used to assess

model reliability more formally, and we will discuss two of these, split-sample analysis and cross-validation more generally, in detail.

## Specifying the Maximum Model

We will use *maximum model* or *full model* to refer to the model with the greatest number of predictors of any model. In model selection, we will consider more parsimonious models than this model. In some cases, models under consideration will be *nested* in the full model, which means that they can be created simply by deleting variables from the full model. In other cases, smaller models may not be nested in the maximum model. A common goal is to find a more parsimonious model that still describes the data (almost) as well as the full model.

When selecting the full model, one must be sure not to select a model that is too large. As a bare minimum, we need $n - p > 0$, where $p$ is the number of columns in $\mathbf{X}$. In order to have fairly stable estimates, we would want $n > 5p$ or $n > 10p$.

Consider a model with 5 covariates, $x_1, x_2, x_3, x_4, x_5$, and an intercept. Several possible "full models" are listed below.

$$
\begin{aligned}
E(\mathbf{y}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 && (1)\\
E(\mathbf{y}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_2^2 \\
&\quad + \beta_7 x_3^2 && (2)\\
E(\mathbf{y}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 && (3)\\
E(\mathbf{y}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 \\
&\quad + \beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5 + \beta_{10} x_1 x_2^2 + \beta_{11} x_1 x_3^2 && (4)\\
E(\mathbf{y}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 \log(x_5) && (5)
\end{aligned}
$$

Suppose that we select Model 2 as our best model. Model 1 would be a candidate model that is *nested* in Model 2, while Model 5 is a *non-nested* candidate model.

## Specifying a Model Selection Criterion

Next, we need to choose a criterion to help us define which model is "best". Some criteria are used to evaluate models with respect to a null or intercept-only model, while other criteria are used to evaluate models with respect to the full model or any larger model (this latter class of criteria is restricted to models *nested* in the full model).

## Comparing a Candidate Model to the Null or Intercept Model

### Adjusted $R^2$

One intuitive model selection criteria is $R^2$. Recall that

$$R^2 = \frac{SSE(\beta_0) - SSE(\beta_0, \ldots, \beta_{p-1})}{SSE(\beta_0)}$$

and provides us with an estimate of the percent of variability explained by the model under consideration. One problem with $R^2$ is that it never decreases when additional variables are added to a model, so that it will always favor selection of the largest model. An *adjusted* $R^2$ is given by

$$R_A^2 = 1 - \frac{n-1}{n-p}(1 - R^2),$$

where $p$ is the number of columns of $\mathbf{X}$. This adjusts the usual $R^2$ for the number of covariates, adding a penalty for models with "too many" covariates. Typically, the model with the largest $R_A^2$ is said to be the best model.

## F Tests

The test of corrected overall regression, given by

$$F_p = \frac{[SSE(\beta_0) - SSE(\beta_0, \ldots, \beta_p)]/(p-1)}{MSE(\beta_0, \ldots, \beta_p)},$$

may be used to test whether the model under consideration offers significant improvement over the intercept-only model. For models without an intercept, the test of uncorrected overall regression may be used to test whether the model under consideration offers significant improvement over the null model.

## Single Model Criteria: AIC and SBC/BIC

The Akaike Information Criterion (AIC) and the Schwarz Criterion or Bayesian Information Criterion (SBC in SAS and BIC many other places) are general metrics. In linear regression,

$$AIC = n \log\left(\frac{SSE}{n}\right) + 2p$$

$$SBC = n \log\left(\frac{SSE}{n}\right) + \log(n)p.$$

For both criteria, smaller is better. Increasing the SSE increases the AIC and SBC, while increasing the number of predictors also increases each measure. AIC tends to favor models that are too large, while the SBC places a greater penalty on larger models. These criteria allow comparison of nested and non-nested models. NOTE: In older versions of SAS, some procedures used "bigger is better" definitions of AIC and SBC, so take care if you are using SAS versions before 8.1.

## Comparing a Candidate Model to the Full Model

### Mallows $C_p$

Mallows $C_p$ compares a candidate model to the full model and is given by

$$C_p = \frac{SSE(\beta_0, \ldots, \beta_{p-1})}{MSE(\text{full})} - n + 2p.$$

Based on the rationale that a model not omitting useful variables should have $SSE(\beta_0, \ldots, \beta_{p-1}) \approx (n-p)\sigma^2$, we seek a model with $C_p \approx p$.

## F Tests

A groupwise test of a candidate model versus the full model, given by

$$F_p = \frac{[SSE(\beta_0, \ldots, \beta_{p-1}) - SSE(\text{full})]/[df(\text{candidate}) - df(\text{full})]}{MSE(\text{full})},$$

may be used to test whether the model under consideration is sufficiently close to the reduced model, provided that the model under consideration is nested in the full model. (Note that this test may be used to test the adequacy of any model nested in a larger model.)

## Specifying the Selection Strategy

### All Possible Regressions

Suppose our full model contains $p$ predictors. An *all possible regressions* strategy involves fitting all possible models. In this case, each of the $p$ predictors (including the intercept) could be in or out of the model, so that we would consider $2^p$ models in our selection process (including a null model). This strategy may get out of hand very quickly, since there are $2^5 = 32$ possible models with 5 columns in $\mathbf{X}$, $2^8 = 256$ possible models with 8 columns in $\mathbf{X}$, and $2^{10} = 1024$ possible models with 10 columns in $\mathbf{X}$. Although SAS will print the "top" models from all possible ones with using some criteria ($R^2$ for example), the time required for an all possible regressions strategy for other criteria (such as $F$ tests against a maximum model) make this strategy infeasible when $p$ is large.

## Backward Elimination

*Backward elimination* begins with the full model and deletes variables with little value. The procedure is described below.

1. Specify the full model and set $p = p_{full}$, the number of columns of the $\mathbf{X}$ matrix in the full model.
2. Fit all $p - 1$ variable models defined by deleting a single variable from the base model. (Typically, we do not consider eliminating the intercept.)
3. For each model, compute an added-last test for the candidate variable.
4. Find the minimum $F$ statistic out of the set of $p - 1$ models (maximum p-value).
   (a) If the corresponding variable is "significant" at a specified level (typically ranging from 0.05 to 0.20), stop and select the model with $p$ predictors as the best model.

(b) If the corresponding variable is not significant, delete the variable in question and set $p = p - 1$.

5. Repeat until a model is chosen.

**Forward Selection**

*Forward selection* begins with the intercept-only (or null) model and adds variables with predictive value. The procedure is described below.

1. Fit the intercept-only (or null) model as the base model so that $p = 1$.
2. Fit all $p + 1$ variable models defined by adding a single variable to the base model.
3. For each model, compute an added-in-order test for the candidate variable.
4. Find the maximum $F$ statistic out of the set of models (minimum p-value).

   (a) If the corresponding variable is "significant" at a specified level

(typically ranging from 0.05 to 0.20), add the variable in question and set $p = p + 1$.

(b) If the corresponding variable is not significant, stop and choose the model with $p$ predictors.

5. Repeat until a model is chosen.

**Stepwise Selection**

Strictly speaking, *stepwise selection* refers to a selection procedure with both forward and backward steps so that addition and deletion of variables may be considered. However, this term is sometimes loosely used to refer to forward selection or backward elimination as well.

SAS implementation: The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the statistic for a variable to be added must be significant at the SLENTRY= level. After a variable is added, however, the stepwise method looks at all

the variables already included in the model and deletes any variable that does not produce an statistic significant at the SLSTAY= level. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model.

Note: It is important to note that these three strategies do not consider all possible models and might actually miss a "best" model. In addition, it is entirely possible that they may select three different models as "best" models. Famous quote from George Box: all models are wrong, but some are useful.

```
data a; do i = 1 to 500;
x1 = 10 + 5*rannor(0); * Normal(10, 25);
x2 = exp(3*rannor(0)); * lognormal;
x3 = 5+10*ranuni(0); * uniform;
x4 = 100 + 50*rannor(0); * Normal(100, 2500);
x5 = x1 + 3*rannor(0); * normal bimodal;
x6 = 2*x2 + ranexp(0);
* lognormal and exponential mixture;
x7 = 0.5*exp(4*rannor(0)); * lognormal;
x8 = 10 +8*ranuni(0); * uniform;
x9 = x2 + x8 + 2*rannor(0);
* lognormal, uniform and normal mix
x10 = 200 +90*rannor(0); * normal(200, 8100);
*x10 = 5*x2 + rannor(0);
y = 3*x2 - 4*x8 + 5*x9 + 3*rannor(0);
* true model with no intercept term;
output; end;
```

```
/*run all possible models*/
proc reg data=a outest=est;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
selection=adjrsq sse aic ;
output out=out p=p r=r; run; quit;

/*run all possible models without intercept*/
proc reg data=a outest=est0;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
noint selection=adjrsq sse aic ;
output out=out0 p=p r=r; run; quit;

/*forward selection*/
proc reg data=a outest=est1;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
slentry=0.15 selection=forward
ss2 sse aic;
```

```
output out=out1 p=p r=r; run; quit;

/*backward selection*/
proc reg data=a outest=est2;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
slstay=0.15 selection=backward
ss2 sse aic;
output out=out1 p=p r=r; run; quit;

/*stepwise selection*/
proc reg data=a outest=est3;
model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 /
slstay=0.15 slentry=0.15 selection=stepwise
ss2 sse aic;
output out=out3 p=p r=r; run; quit;

/*stepwise group selection*/
```

```
proc reg data=a outest=est4;
model y={x1 x2} x3 x4 x5 x6 x7 x8 x9 x10 /
selection=stepwise slstay=0.15 slentry=0.15
groupnames=x1 x2 x3 x4 x5 x6
x7 x8 x9 x10;
```

## Comments

1. In any selection strategy, one must take steps to ensure that nonsensical models do not emerge as "winners". For example, if $x, x^2$, and $x^3$ are in the full model, then we do not wish to allow a backward selection strategy to delete $x$ while letting $x^2$ and $x^3$ remain in the model. In such cases, we will define variable "groups" to be tested as units. In this case, a logical strategy would be to treat $(x, x^2, x^3)$ as a group. In backward elimination, we would first test the group. If it was significant, we might then move to test $x^3$, and if significant, we would retain all three variables in the model.

2. The p-values obtained using a series of $F$ tests should not be viewed strictly but merely as guides due to inflated type I error rates due to multiple testing.

3. Ideally, the best error term is the one from the full model because it should have the least bias. In forward selection in particular, the

estimate of the error term may be biased upwards, leading the process to select a final model that is too small.

4. A group-wise, backward selection strategy based on $F$ tests is superior in a wide variety of settings.

## Fixed Tests Model Selection Strategy

If one wishes to avoid "data dredging" and to better preserve the type I error rate, a fixed tests implementation should be considered. This strategy is outlined below.

1. Very thoughtfully specify the maximum model.
2. Select groups and use a groupwise strategy, selecting a sequence for testing models and adjusting $\alpha$ for tests.
3. Assess diagnostics.
   (a) Fit the full model and check for collinearity.
   (b) Parsimoniously change or reduce the full model to eliminate collinearity.
   (c) Conduct assumption diagnostics.
4. Conduct the *planned* tests.
5. Conduct assumption diagnostics on the resulting reduced model.

## Assessing Reliability

Why is assessing reliability an important step in model fitting?

Consider the following data.

One possible model for the data is a simple linear regression, shown below.

Another approach is a regression model with linear and quadratic terms in $x$.

A third approach is a linear nonparametric approach ("connect the dots").

Why not choose the model with the best fit to the data?

How well will the model predict future data from the same distribution?

## Split-Sample Analysis

The simplest model validation method is one-time *data-splitting*. In this method, the data are split into *training* (model development) and *test* (model validation) samples by a random process. This splitting must occur before any analysis begins. Then, regression is performed on the training set, with future performance evaluated in the test set.

The split-sample procedure is outlined below.

1. Before looking at the data, split the sample using a random process (this process may be stratified by subgroups defined by values of $\mathbf{X}$ such as smoking status or race). The split fraction could be 50-50(so that half the data fall into the training sample and half fall into the test sample). For large data sets, often a fraction as small as 10% could be placed in the training sample while still providing a sufficient sample size for exploratory data analysis. If the sample size is too small to allow a reasonable fraction in the training sample, the value of a split-sample

approach will be greatly diminished.

2. Conduct any and all desired exploratory analyses on the training data, including diagnostics.

3. Based on the exploratory analyses, select a "best" model based on selected criteria. The parameter estimates for this model will be called $\widehat{\boldsymbol{\beta}}_1$.

4. Compute the predicted values $\widehat{\mathbf{y}}$ for this model and call them $\widehat{\mathbf{y}}_1$.

5. Compute the squared multiple correlation, $R_1^2 = r^2(\mathbf{y}_1, \widehat{\mathbf{y}}_1)$ for this model.

6. Compute the cross-validation correlation.

   (a) Compute the predicted values in the test sample as $\widehat{\mathbf{y}}_2 = \mathbf{X}_2\widehat{\boldsymbol{\beta}}_1$. (That is, use the $\widehat{\boldsymbol{\beta}}$ from the training data with the covariates from the test data to compute predicted values.)

   (b) Compute the squared cross-validation correlation as $R_{*2}^2 = r^2(\mathbf{y}_2, \widehat{\mathbf{y}}_2)$.

(c) Compute the estimated shrinkage on cross-validation as $R_1^2 - R_{*2}^2$. Smaller shrinkages are better. (One rule of thumb is that shrinkage $<$0.05 indicates results are generalizable and that shrinkage $>$ 0.10 is cause for concern.) Often, one computes the proportion relative shrinkage by dividing by $R_1^2$.

(d) The MSE may also be used to evaluate future performance.

1. Conduct regression diagnostics on the pooled data.
   (a) If shrinkage is "small enough," pool the data to provide best available estimates of $\boldsymbol{\beta}$.
   (b) If shrinkage is "poor," pool data to conduct a second-round *exploratory* analysis.
2. Report results honestly.

One big advantage of data-splitting is that hypothesis tests are confirmed in the test sample. In addition, this method is simple to implement. However, it has several disadvantages, including the following.

- Data-splitting greatly reduces the sample size for both model development and model testing.
- Different splits may lead to different results, and with small sample sizes, our test set may just be lucky or unlucky.
- Data-splitting does not validate the final model but rather a model

developed on only a subset of the data. The training and test sets are recombined for fitting the final model, which is not validated.

## Cross-Validation

*Cross-validation* is a generalization of data-splitting that overcomes some of its disadvantages.

*Leave-out-one cross-validation* is similar to jackknifing. In leave-out-one cross-validation, one observation is omitted from the analysis, and the response for that one subject is predicted using a model obtained from the remaining $n - 1$ subjects. Note the error of that prediction. When you've done this $n$ times (once for each subject), report the mean squared error over all the predictions. This method does not "waste" data, but is computationally expensive.

Consider leave-out-one cross-validation for the previous data. Omitting the $j^{th}$ observation, $j = 1, \ldots, 10$, we have the lines below for the linear models.

The errors are calculated as the distance from the fitted line omitting each subject to each subject's observed outcome. Summing these squared distances, we obtain $SSE$. The estimated error from the linear model can be compared to that of the quadratic and "connect-the-dots" models to see which model is preferred.

*K-fold cross validation* randomly breaks the dataset into $k$ partitions. For each partition, train on the points not in the partition, and find the test-set errors for the partition of interest. Repeat for all $k$ partitions, and then report the mean squared error over all the predictions. This type of cross-validation (often done for $k = 10$) is a compromise between the leave-out-one and the split sample approaches.

## Final Comments

Model selection and validation tend to be personal topics with different investigators preferring a variety of different methods. The major consensus is that exploratory analyses must be treated with caution, and test results must be presented honestly (and not as "confirmatory" when extensive model exploration has been carried out).

Frank Harrells comments:
Here are some of the problems with stepwise variable selection.

- It yields R-squared values that are badly biased to be high.
- The F and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
- The method yields confidence intervals for effects and predicted values that are falsely narrow.
- It yields p-values that do not have the proper meaning, and the proper correction for them is a difficult problem.

- It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large).
- It has severe problems in the presence of collinearity.
- It is based on methods (e.g., F tests for nested models) that were intended to be used to test prespecified hypotheses.
- Increasing the sample size doesn't help very much.

For Post-Selection Statistical Inference, you may check Robert Tibshirani's talk.

## Coding Schemes for Regression

*Reading Assignment:*

- Muller and Fetterman, Chapter 12: "Coding Schemes for Regression" (Required)

*Goals*

1. Understand various coding schemes for ANOVA and relationships

between ANOVA and multiple regression models.

2. Understand basic strategy and techniques of multiple comparisons.

3. Recognize the impact of various amounts of missing data.

The study of ANOVA is motivated by desire to model and test hypotheses about two or more group means.

## Regression or ANOVA?

In theory and computation, ANOVA is a special case of regression analysis with dummy variables as predictors.

An *indicator* or *dummy* variable is used to represent group membership. Suppose we are interested in the effect of two weight loss regimens, diet and exercise, over a period of two months, compared to subjects on neither regimen (control subjects). Then we can create three dummy variables to denote regimen membership as follows.

$$
\begin{aligned}
x_1 &= \begin{cases} 1 & \text{diet} \\ 0 & \text{exercise or neither} \end{cases} \\
x_2 &= \begin{cases} 1 & \text{exercise} \\ 0 & \text{diet or neither} \end{cases} \\
x_3 &= \begin{cases} 1 & \text{neither} \\ 0 & \text{diet or exercise} \end{cases}
\end{aligned}
$$

For analysis, we will use $(x_1, x_2, x_3)$ (or some combination of them) in our $\mathbf{X}$ matrix to represent the diet regimens.

The name ANOVA reflects the expression of tests in terms of variances and the nature of computational strategies predating computers.

An ANOVA *coding scheme* defines a set of rules for representing all of the information in one or more categorical variables as one or more interval variables. We will discuss the following five coding schemes:

1. reference cell,
2. cell mean,
3. classical ANOVA,
4. effect, and
5. polynomial.

Each coding scheme merits consideration as well as a description of its relationship to other schemes. This chapter only considers estimation, and all consideration of testing will be left to subsequent chapters.

## Classical ANOVA Coding

*Classical ANOVA coding* uses a less than full rank $\mathbf{X}$ matrix (rank$(\mathbf{X}_{n\times(p+1)}) = p$) intentionally. This $\mathbf{X}$ matrix is equal to the cell mean coding $\mathbf{X}$ matrix with the addition of a column of $1$'s for the

intercept. The model is given by

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\,\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $\mathbf{X}$ has rank $p$ (less than full rank).

Often, we represent $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ as $\boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_p)$.

## Expected Values for Classical ANOVA Coding

| Group | $E[y_i]$ | Group Mean |
|---|---|---|
| 1 | $1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$ | $\mu_1$ |
| 2 | $1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_2$ | $\mu_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| p | $1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_p = \beta_0 + \beta_p$ | $\mu_p$ |

Group 1 is the reference group, and $\beta_0 + \beta_1 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_2) - (\beta_0 + \beta_1)) = (\beta_2 - \beta_1)$, and the difference in mean response between group 2 and group $p$ is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_2) - (\beta_0 + \beta_p)) = (\beta_2 - \beta_p)$.

Because $\mathbf{X}$ is not full rank, our estimates of $\boldsymbol{\beta}$ are not unique because $\boldsymbol{\beta}$ is not estimable. However, we may obtain a unique solution by imposing an additional constraint. A common constraint is to require $\sum_{j=1}^{p} \beta_j = 0$. This ensures that $\beta_0$ is the grand mean because

$$\begin{aligned} \frac{(\beta_0 + \beta_1) + (\beta_0 + \beta_2) + \cdots + (\beta_0 + \beta_p)}{p} &= \frac{p\beta_0 + \sum_{j=1}^{p} \beta_j}{p} \\ &= \beta_0. \end{aligned}$$

Many authors describe $\beta_0 = \mu$ as the grand mean.

Classical ANOVA coding leads to a less than full rank model that may be numerically unstable. Because all parameters are not estimable, we must use the theory for the less than full rank model. However, it is easier just to use a full rank coding scheme.

SAS uses classical ANOVA coding but imposes the constraint that $\beta_j = 0$ for one $j$ (a *reference cell* constraint).

Note that parameter estimates under different constrains may have different meanings. For example, $\widehat{\beta}_0$ under the first constraint is the grand mean, while $\widehat{\beta}_0$ under the second constraint is the mean for the reference group.

## Reference Cell Coding

### One Group

Suppose that a geneticist wishes to address the impact of a potentially toxic environmental exposure on mice. The intercept-only model assumes that all responses differ randomly from a common mean response, called the *grand mean*. Thus we have the model

$$\begin{aligned} \mathbf{y}_{n\times 1} &= \mathbf{X}_{n\times 1}\,\boldsymbol{\beta}_{1\times 1} + \boldsymbol{\varepsilon}_{n\times 1} \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \beta_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \end{aligned}$$

This ANOVA model is often called the *grand mean model*. In this model, $\mathbf{X}$ is full rank, and $E[y_i] = \beta_0$. Often, we replace $\beta_0$ with $\mu$ so that $E[y_i] = \mu$.

In this model,

$$\begin{aligned}\widehat{\beta_0} &= (\mathbf{J}_n'\mathbf{J}_n)^{-1}\mathbf{J}_n'\mathbf{y} \\ &= n^{-1}\sum_{i=1}^{n} y_i \\ &= \overline{y}.\end{aligned}$$

## Two Groups

Now suppose that the mice are from different mouse strains so that $n_1$ of the mice are black six mice, and $n - n_1 = n_2$ of the mice are Swiss

albino mice. Then we have the model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times 2}\,\boldsymbol{\beta}_{2\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{J}_{n_1\times 1} & \mathbf{0}_{\mathbf{n_1}\times \mathbf{1}} \\ \mathbf{J}_{n_2\times 1} & \mathbf{J}_{n_2\times 1} \end{bmatrix}_{n\times 2}.$$

A dummy (indicator) variable indicates the species for each

observation; that is,

$$x_1 = \begin{cases} 0, & \text{black six} \\ 1, & \text{Swiss albino} \end{cases}.$$

Predictor values assigned by the scientist are often called treatment levels.

**Three or More Groups**

For $p$ mouse species, we have the model

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $\mathbf{X}$ is full rank $p$.

Expected Values for Reference Cell Coding

| Group | $E[y_i]$ | Group Mean |
|---|---|---|
| 1 | $1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0$ | $\mu_1$ |
| 2 | $1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$ | $\mu_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| p | $1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_0 + \beta_{p-1}$ | $\mu_p$ |

Group 1 is the reference group, and $\beta_0 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_1) - \beta_0) = \beta_1$, and the difference in mean response between group 2 and group $p$ is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_1) - (\beta_0 + \beta_{p-1})) = (\beta_1 - \beta_{p-1})$.

## Cell Mean Coding

With *cell mean coding*, all of the $\beta$'s equal group means. Cell mean coding is the most natural coding scheme and the easiest to understand (and the easiest to explain to investigators!). For $p$ groups, we create $p$ indicator variables and do not include an intercept in the

model, which is given by

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $\mathbf{X}$ is full rank $p$.

Expected Values for Cell Mean Coding

| Group | $E[y_i]$ | Group Mean |
|---|---|---|
| 1 | $1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0$ | $\mu_1$ |
| 2 | $0 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_1$ | $\mu_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| p | $0 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_{p-1}$ | $\mu_p$ |

The vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ is often written $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_p)$ or $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_p)$.

Group 1 is the reference group, and $\beta_0 = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = (\beta_1 - \beta_0)$, and the difference in mean response between group 2 and group $p$ is $(\mu_2 - \mu_p) = (\beta_1 - \beta_{p-1})$. Does this model span an intercept?

## Effect Coding

*Effect coding* provides a useful coding scheme. This full rank scheme has design matrix $\mathbf{X}$ with a column of 1's and $p-1$ other columns, like the reference cell coding scheme. However, the *effect coding scheme* reassigns the value of 0 to -1 for dummy variable covariates for

subjects in the reference group. The model is given by

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & & & & \vdots \\ 1 & -1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $\mathbf{X}$ is full rank $p$.

The effect coding indicator variable may be created in two steps.

1. Create a dummy variable indicating group membership, with the value $1$ if the observation belongs to the group and $0$ otherwise.
2. If the observation represents the reference group, then reassign the value of $0$ to $-1$.

Expected Values for Effect Coding

| Group | $E[y_i]$ | Group Mean |
|---|---|---|
| 1 | $1 \cdot \beta_0 - 1 \cdot \beta_1 - \cdots - 1 \cdot \beta_{p-1} = \beta_0 - \sum_{j=1}^{p-1} \beta_j$ | $\mu_1$ |
| 2 | $1 \cdot \beta_0 + 1 \cdot \beta_1 + \cdots + 0 \cdot \beta_{p-1} = \beta_0 + \beta_1$ | $\mu_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| p | $1 \cdot \beta_0 + 0 \cdot \beta_1 + \cdots + 1 \cdot \beta_{p-1} = \beta_0 + \beta_{p-1}$ | $\mu_p$ |

Group 1 is the reference group, and $\beta_0 - \sum_{j=1}^{p-1} \beta_j = \mu_1$ is the mean for group 1. The difference in mean response between group 2 and group 1 is $(\mu_2 - \mu_1) = ((\beta_0 + \beta_1) - (\beta_0 - \sum_{j=1}^{p-1} \beta_j) = 2\beta_1 + \sum_{j=2}^{p-1} \beta_j$, and the difference in mean response between group 2 and group $p$ is $(\mu_2 - \mu_p) = ((\beta_0 + \beta_1) - (\beta_0 + \beta_{p-1})) = (\beta_1 - \beta_{p-1})$.

Note that the mean of all group means is

$$\frac{(\beta_0 - \sum_{j=1}^{p-1} \beta_j) + (\beta_0 + \beta_1) + \cdots + (\beta_0 + \beta_p)}{p} = \frac{p\beta_0}{p} = \beta_0.$$

This parameter is the mean of the particular cells in the current design. This is different from the grand mean parameter in classical ANOVA coding (when there is no restriction), which represents the hypothetical mean of all observations in the population.

## Relationships Among Coding Schemes

Any full rank $\boldsymbol{X}$ may be expressed as a full rank linear transform of any other (with both based on the same categorical predictors). Any parameter estimable or testable in one coding scheme is also estimable or testable in any other.

With $p$ groups, only $p$ parameters are estimable.

Parameters for any coding scheme may be expressed as linear functions of cell means.

**Next: One-way ANOVA**

*Reading Assignment:*

- Muller and Fetterman, Chapter 13: "One-Way ANOVA"

# Lecture 16: One-Way ANOVA

*Reading Assignment:*

- Muller and Fetterman, Chapter 13: "One-Way ANOVA"

We use analysis of variance (ANOVA) to answer questions like the following.

- Do two or more groups differ in mean response?
- Does the new drug reduce symptoms when compared to existing drugs?
- Does group membership differentially predict response?

Now we shift from considering continuous predictors to considering categorical predictors. ANOVA models are a type of GLM used with

categorical predictors.

## Specification of the Model

We want to determine whether two or more groups differ in location. The Gaussian errors assumption reduces this task to testing equality of means, generalizing Student's T test to three or more groups.

**Example: Testing Equality of Cell Means**

Pagano and Gauvreau (1993) present lung function data for heart patients at three medical centers: Johns Hopkins University, Rancho Los Amigos, and St. Louis University. For each center, investigators measured lung function as $FEV_1$, the forced expiratory volume in 1 second in liters.

Below is a plot of the $FEV_1$ for each subject by center.

The $FEV_1$ of patients across centers does differ, but there is also a lot of variation within center. While Rancho Los Amigos does have the highest lung function on average, some patients at Johns Hopkins have higher $FEV_1$ than some of the Rancho Los Amigos patients, for example.

We need a statistical test to determine whether this is a statistically significant difference. In ANOVA (as well as regression more generally), we ask questions about *means* of groups by analyzing their *variances*...how does this work?

## GLM Assumptions in ANOVA

| Assumption | ANOVA Interpretation |
|---|---|
| H | within cell / between cell |
| I | check sampling scheme |
| L | automatically OK, given design |
| E | no problem in practice |
| Gauss | within cell |

Violation of the independence assumption may inflate $\alpha$ greatly. We must verify independence with careful thought about the sampling scheme, e.g., one individual may be sampled twice. To evaluate the assumption of Gaussian distribution of residuals, we may look at the Gaussian errors assumption within each group separately if the sample size in each group is large enough ($\geq 30$ say). Otherwise, we may look at studentized residuals pooled over all the groups.

## The Primacy of Cell Means

Cell mean coding will provide our default choice, reflecting our primary interest in comparing means across groups. Recall that for cell mean coding with $G$ groups, we create $G$ indicator variables and do not include an intercept in the model, which is given by

$$
\begin{aligned}
\mathbf{y}_{n\times 1} &= \mathbf{X}_{n\times G}\,\boldsymbol{\beta}_{G\times 1} + \boldsymbol{\varepsilon}_{n\times 1} \\
\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{G-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix},
\end{aligned}
$$

where $\mathbf{X}$ is full rank $G$. Matrix $\mathbf{X}$ can be summarized with the essence matrix, the matrix created by deleting any duplicate rows from the design matrix. The essence matrix helps us to determine rank and the

relationships between parameter definitions. What is the essence matrix of $\mathbf{X}$? (answer: $\text{Es}(\mathbf{X}) = \mathbf{I}$).

Often, double indices are used to denote $y$'s by group as follows.

$$\begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{Gn_G} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 1 & 0 & \cdots & \cdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 1 & 0 & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \vdots \\ \beta_{G-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Parameters in other codings are linear transformations of cell means.

Consider cell mean coding for $G$ groups, with $n = \sum_{g=1}^{G} n_g$. Then

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n_1 & & & 0 \\ & n_2 & & \\ 0 & & \ddots & \\ & & & n_G \end{bmatrix}_{G \times G}$$

$$\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \sum_{j=1}^{n_1} y_{1j} \\ \sum_{n=1}^{n_2} y_{2j} \\ \vdots \\ \sum_{n=1}^{n_G} y_{Gj} \end{bmatrix}_{G \times 1}$$

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \overline{y}_{1\cdot} \\ \\ \overline{y}_{2\cdot} \\ \\ \vdots \\ \\ \overline{y}_{G\cdot} \end{bmatrix}_{G\times 1}$$

Note that in *cell mean* coding, our estimates are the sample means for each group! In other coding schemes, estimates are functions of cell means.

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \mathbf{J}\overline{y}_{1\cdot} \\ \mathbf{J}\overline{y}_{2\cdot} \\ \vdots \\ \mathbf{J}\overline{y}_{G\cdot} \end{bmatrix}_{n\times 1} \qquad \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \begin{bmatrix} y_{11} - \overline{y}_{1\cdot} \\ \vdots \\ y_{1n_1} - \overline{y}_{1\cdot} \\ y_{21} - \overline{y}_{2\cdot} \\ \vdots \\ y_{Gn_G} - \overline{y}_{G\cdot} \end{bmatrix}_{n\times 1}$$

$$\widehat{\sigma}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{(n-G)} = \frac{SSE}{(n-G)} = \frac{\sum_{i=1}^{G}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2}{n-G}$$

## (Usual) Overall Test

We wish to test the hypothesis that the means are the same for each group. For three groups, we have

$$H_0: \quad \beta_0 = \beta_1 = \beta_2 \qquad \text{(cell mean)}$$

$$\Leftrightarrow$$

$$H_0: \quad \beta_1 = \beta_2 = 0 \qquad \text{(reference cell)}$$

$$\Leftrightarrow$$

$$H_0: \quad \beta_1 = \beta_2 = 0 \qquad \text{(effect)}$$

The overall test involves equality of $G$ means, which implies $G-1$ constraints so that $\boldsymbol{\theta}$ is $(G-1) \times 1$ and $a = G-1$.

We are testing $H_0$: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

We state $\boldsymbol{C}$ (unique to coding scheme), use $\boldsymbol{\theta}_0 = \mathbf{0}$, and then compute $SSH = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \boldsymbol{M}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ and $f_{obs} = (SSH/a)/\widehat{\sigma}^2$.

## Reference Cell Coding

Recall that for reference cell coding with three groups, we have the model

$$\mathbf{y} = \mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\text{Es}(\mathbf{X}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. To test the hypothesis of equal cell means, we have $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ implying that $\boldsymbol{\theta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. We use $\boldsymbol{\theta}_0 = \mathbf{0}$.

## Effect Coding

Recall that for effect coding with three groups, we have the model

$$\mathbf{y} = \mathbf{X} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\text{Es}(\mathbf{X}) = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. To test the hypothesis of equal cell means, we have $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ implying that $\boldsymbol{\theta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$. We use $\boldsymbol{\theta}_0 = \mathbf{0}$.

If any $G - 1$ pairwise differences are zero, then all pairwise differences are zero, so the traditional overall test necessarily has $G - 1 = a$ degrees of freedom. All the above coding schemes give the same

p-value, $F$, and $R^2$ because they are full rank transformations of the rows; e.g.,$\boldsymbol{C}_1 = \boldsymbol{T}\boldsymbol{C}_2$.

**Traditional Source Table**

Traditional ANOVA Source Table

| Source | $\boldsymbol{df}$ | $\boldsymbol{SS}$ | Mean Square | $\boldsymbol{F_{obs}}$ | $\boldsymbol{p}$ |
|---|---|---|---|---|---|
| Between | $G-1$ | | $SSH/a$ | | |
| Within | $n-G$ | | $\widehat{\sigma}^2$ | | |
| (Error) | | | | | |
| Total | $n-1$ | | | | |
| (Corrected) | | | | | |

## Estimating and Testing Cell Means in One-Way ANOVA

Using the essence matrix as the contrast matrix defines the cell means. For cell mean coding, the parameter estimates are the cell means. For other coding schemes, we use the parameter estimates provided to calculate the cell means.

Once you determine the cell means for each coding scheme, it is straightforward to test contrasts in the cell means.

## Which Means are Different?

If we reject the null hypothesis that all means are equal, naturally we next want to know which means are in fact different.

There are $\binom{G}{2} = \frac{G(G-1)}{2}$ pairwise comparisons, and infinitely many others (if three or more means are involved).

An overall hypothesis that all means are equal tests all linear combinations of means and not just pairwise comparisons.

For the overall test with cell mean coding,

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

creates a matrix of secondary parameters which are mean differences. Each row of $\boldsymbol{C}$ defines a contrast matrix, a GLH, and an $F$ test. Any

$\boldsymbol{C}$ with a single row creates an $F$ test with one numerator $df$ and hence corresponds to a $T$.

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix} \qquad H_0\colon\ \beta_0 = \beta_1$$
$$\boldsymbol{C} = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \qquad H_0\colon\ \beta_0 = \beta_2$$
$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \qquad H_0\colon\ \beta_1 = \beta_2$$

## Contrasts

Describe a linear combination of cell means, $\sum_{i=1}^{G} c_i \mu_i$, as a **contrast** if $\sum_{i=1}^{G} c_i = 0$. Pairwise contrasts (pairwise comparisons) use coefficients such as $\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \end{bmatrix}$ or $\boldsymbol{C} = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix}$.

Complex contrasts, such as $\boldsymbol{C} =$

$$\begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

involve 3 or more means.

The null and alternative hypotheses in this case are

$$H_0: \qquad \sum_{i=1}^{G} c_i \mu_i = 0$$

$$H_A: \qquad \sum_{i=1}^{G} c_i \mu_i \neq 0.$$

A planned (*a priori*) contrast requires choosing the number of contrasts and specifying the associated $\boldsymbol{C}$ matrices before seeing any data. Unplanned (*a posteriori*) contrasts include all others not specified in advance.

For **balanced** data, a set of contrasts is **orthogonal** if $\boldsymbol{C}\boldsymbol{C}'$ is diagonal (rows orthogonal).

## Example: $FEV_1$ Data

The following SAS code creates the design matrix for cell mean coding and conducts the overall test of equality of group means for the lung function data.

```
data fev2;
set fev;
if center="hopkins" then h_ind=1; else h_ind=0;
if center="ranchola" then r_ind=1; else r_ind=0;
if center="stlouis" then s_ind=1; else s_ind=0;
run;

proc glm data=fev2;
model fev=h_ind r_ind s_ind/noint;
contrast Usual Overall Test h_ind 1 r_ind -1 s_ind  0,
                            h_ind 1 r_ind 0  s_ind -1;
run;
```

The results of the test are provided below.

```
                              The GLM Procedure

                            Dependent Variable: fev
```

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 482.5697439 | 160.8565813 | 633.19 | <.0001 |
| Error | 57 | 14.4802561 | 0.2540396 | | |
| Uncorrected Total | 60 | 497.0500000 | | | |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| h_ind | 1 | 144.8344048 | 144.8344048 | 570.13 | <.0001 |
| r_ind | 1 | 147.1369000 | 147.1369000 | 579.19 | <.0001 |
| s_ind | 1 | 190.5984391 | 190.5984391 | 750.27 | <.0001 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| h_ind | 1 | 144.8344048 | 144.8344048 | 570.13 | <.0001 |
| r_ind | 1 | 147.1369000 | 147.1369000 | 579.19 | <.0001 |
| s_ind | 1 | 190.5984391 | 190.5984391 | 750.27 | <.0001 |

| Contrast | DF | Contrast SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|

```
Usual Overall Test  2      1.58283723      0.79141861       3.12    0.0520


                                   Standard
Parameter         Estimate            Error    t Value    Pr > |t|

h_ind          2.626190476      0.10998692      23.88      <.0001
r_ind          3.032500000      0.12600585      24.07      <.0001
s_ind          2.878695652      0.10509614      27.39      <.0001
```

We see that there is a marginally significant difference among the cell means ($F = 3.12$ with 2 numerator and 57 denominator degrees of freedom yields $p = 0.05$). At least one center has average $FEV$ that is different from the others.

What is the interpretation of the overall F test and the nine tests for the individual groups?

## Step-Down Testing

Describe a test logically subsumed by another test as a **step-down test**. For example, having conducted an (overall) test for equality among a set of means, one may wish to step down to test equality of two particular means. In the population, the truth of the null hypothesis of the parent implies the truth of the null hypothesis of the child, and the truth of the alternative hypothesis for the child implies the truth of the alternative hypothesis for the parent.

Any linear combination of the rows of $\boldsymbol{C}$ represents a step-down test.

## Properties of Cell Mean Estimates

If $\boldsymbol{C}$ and $\boldsymbol{X}$ are both orthogonal, then the contrast estimates $\widehat{\theta}_k$ are uncorrelated and statistically independent (under HILE Gauss). This occurs when we estimate cell means with cell mean coding.

There are at most $G-1$ mutually orthogonal contrasts (of mean differences) but infinitely many sets of contrasts are possible. There can be $G$ mutually orthogonal contrasts if the set spans the grand mean.Recall that

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &\sim \mathcal{N}_a[\boldsymbol{\theta},\, \sigma^2\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}'] \\ &\sim \mathcal{N}_a[\boldsymbol{\theta},\ \sigma^2\boldsymbol{M} \qquad\qquad ]\end{aligned}$$

Each row of $\boldsymbol{C}$ defines a $\theta_k$, with $\widehat{\theta}_k \sim \mathcal{N}(\theta_k, \sigma^2 m_{kk})$ and

$$\frac{\widehat{\theta}_k - \theta_0}{\sqrt{\widehat{\sigma}^2 m_{kk}}} \sim T\left(n-r\right),$$

where $r$ is the rank of $\mathbf{X}$.

For cell mean coding the $\widehat{\theta}_k$'s are independent for orthogonal $\boldsymbol{C}$.

$\widehat{\theta}_k$ tests are never exactly independent (for finite $n$) because they all use $\widehat{\sigma}^2$.

We invert the test to create confidence intervals:

$$\widehat{\theta}_k \pm \ (\widehat{\sigma} \cdot \sqrt{m_{kk}}) \cdot t_{crit}$$

with $t_{crit} = F_T^{-1}(1 - \alpha/2; n - r)$

## Conducting Multiple Comparisons

### Choosing An Error Rate

The type I error rate for a collection of tests may be higher than $\alpha$.

$K$ independent tests with a nominal type I error rate of $\alpha$ lead to a probability of one or more false positives of $1-(1-\alpha)^K$. Using $\alpha = .05$ and $K = 10$ implies $1-(1-\alpha)^K \approx 0.40 >> 0.05$.

For dependent tests, the Bonferroni inequality provides an upper bound of $K\alpha$, which equals $0.50$ in this example. For the Bonferroni approach, test each comparison at $\alpha_k = \alpha/K = 0.05/10 = 0.005$.

## Overview of Methods for Multiple Comparisons

Only a few of many techniques for controlling the error rate will be highlighted here.

The Bonferroni approach is completely general for planned (*a priori*) comparisons.

The development of alternative methods arose from a desire for

1. increased statistical power,
2. the ability to control unplanned comparisons, or
3. increased robustness to violation of GLM assumptions.

For a small number of planned contrasts, use the Bonferroni correction.

With an overall test size of $.01$ and four contrasts, $\alpha_k = \alpha/K$ implies a nominal size for each test of $.01/4 = .0025$.

Some authors recommend not using any correction in the special case of a set of planned and orthogonal contrasts because a significant overall test implies a trend of some sort, and the step-down tests merely need to identify the order of the trend. This approach reduces the chances of having a significant overall test with no particular trend significant.

Dunnett's (1955, 1964) test was designed to test each of $G-1$ treatments against a control. Considering any other comparison with Dunnett's correction inflates test size.

Tukey's HSD (honest significant difference) method and generalizations of it (Tukey, 1953; Kramer, 1956) allow testing all pairwise comparisons.

Scheffé's (1953, 1959) test controls test size with any contrast, including complex contrasts (non-pairwise), whether planned or not. This test also protects against testing infinitely possible contrasts. It

has the desirable property of *coherence*. That is, the failure of the overall test ensures no significant step-down tests (e.g.,mean comparison tests) exist, while a significant overall test ensures at least one significant step-down test.

For a small number of comparisons, a Bonferroni correction suffices. For a larger number or for unplanned comparisons, Scheffé's test is preferred.

## Example Formulas

In general, $\widehat{\boldsymbol{\theta}} \sim \mathcal{N}_a(\boldsymbol{\theta}, \sigma^2 \boldsymbol{M})$, with $\boldsymbol{M} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}'$.

For a cell mean coding with $G$ groups, $(\boldsymbol{X}'\boldsymbol{X}) =$Dg$(n_1, n_2, \ldots n_G)$, which implies $(\boldsymbol{X}'\boldsymbol{X})^{-} =$Dg$(1/n_1, 1/n_2, \ldots 1/n_G)$. For $\boldsymbol{C}_k =$row$_k(\boldsymbol{C})$ and *cell mean coding*,

$$m_{kk} = \boldsymbol{C}_k(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}_k' = \sum_{g=1}^{G} \frac{c_{kg}^2}{n_g}.$$

A Bonferroni correction for planned comparisons may be implemented with any linear models program merely by proper specification of the test size.

For other corrections, most of the work involves selecting the proper critical value based on determining the set of comparisons.

Scheffé's method for unplanned comparisons provides the most general and sturdy technique. Its downside is a lack of power.

## Example: Testing Combinations of Means for FEV Data

We use the following SAS code to test all possible differences among cell means for the FEV data, using Scheffé's correction.

```
proc glm data=fev2;
class center;
model fev=center/noint;
lsmeans center/ pdiff adjust=scheffe;
run;
```

The LSMEANS statement produces the least squares estimates of CLASS variable means, where CLASS is used to indicate to SAS which variables are categorical. You may user the ORDER= option in the proc glm statement to alter how SAS chooses the reference level in proc glm (several options). In one-way ANOVA, these are just the usual means. In higher-way ANOVA models, the least squares means are simply the means of group means (thus if the design is not *balanced*, LSMEANS are not simple means).

This code yields the following additional output.

```
              Least Squares Means
   Adjustment for Multiple Comparisons: Scheffe

                                      LSMEAN
     center         fev LSMEAN        Number
    hopkins        2.62619048            1
    ranchola       3.03250000            2
    stlouis        2.87869565            3

      Least Squares Means for effect center
       Pr > |t| for H0: LSMean(i)=LSMean(j)

              Dependent Variable: fev

   i/j              1             2             3
     1                       0.0603        0.2605
     2         0.0603                      0.6466
     3         0.2605        0.6466
```

We see some evidence that the mean $FEV_1$ at Rancho Los Amigos (3.03 liters) is higher than the mean $FEV_1$ at Johns Hopkins (2.63 liters), but the other means do not appear to differ significantly.

## Trend Tests

Usually, orthogonal sets of tests are unappealing because $G-1$ tests do not cover all interesting mean comparisons.

For example, if the categorical predictor can be interpreted as an interval scale variable (e.g.,low, medium, or high dose), then the $G-1$ orthogonal trend tests are appealing.

Assume we are interested in one species of tree, grown at 200, 400, or 600m, with corresponding hypothesis $H_0$: $\mu_{200} = \mu_{400} = \mu_{600}$. Having rejected $H_0$, what is trend across elevation?

First, we can plot the observed treatment means as a function of the factor levels. This plot may provide a general idea of any pattern that is present.

$G$ levels imply $G-1$ trends can be tested. With equal spacing and equal cell sizes, we may use the orthogonal polynomial coefficients in

Table B-10. Without these conditions, tests are more complicated.

Using the contrast matrices described here with cell mean coding allows computing exactly the same tests as available in polynomial coding, which equal tests in polynomial regression.

For *cell mean* coding, consider

$$\boldsymbol{C}_{\mathsf{O}} = \begin{bmatrix} -1 & 0 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{matrix} \text{linear} \\ \text{quadratic} \end{matrix}$$

This spans the matrix commonly used for the overall test, namely

$$\boldsymbol{C} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

in that each equals a full rank transformation of the rows of the other ($\boldsymbol{C}_{\mathsf{O}} = \boldsymbol{T}\boldsymbol{C}$, with $\boldsymbol{T}$ a full rank matrix of dimension $a \times a$).

The GLH p-value and test statistic do not change with such a transformation.

Plotting the elements of a row of $\boldsymbol{C}_{\mathsf{O}}$ against the column # displays the shape indicated by the labeling.

Separate one $df$ tests provide a set of $G-1$ orthogonal trend tests.

It seems reasonable to ignore any Bonferroni or other sort of correction in this setting if the trend tests represent the only planned comparisons.

## Homogeneity

There are only $G$ distinct groups, so comparing variances for $G$ cells is useful for assessing homogeneity of variances. The Hartley's test involves computing the ratio of the largest group variance, max($s_j^2$) to the smallest group variance, min($s_j^2$). The resulting ratio, $F_{\max}$, is then compared to a critical value from a table of the sampling distribution of $F_{\max}$. If the computed ratio is less than the critical value, the groups are assumed to have similar or equal variances.

Bartlett's test is used to test the null hypothesis, $\mathrm{H}_0$ that all $k$ population variances are equal against the alternative that at least two are different. It is a modification of the normal-theory likelihood ratio test. Bartlett's test is sensitive to departures from normality. That is, if the samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality.

Newer tests (the Brown and Forsythe test or O'Brien test)(HOVTEST=BF or HOVTEST=OBRIEN in PROC GLM) are much more robust to the underlying distribution. They transform the original values of the dependent variable to derive a dispersion variable and then to perform analysis of variance on this variable.

The significance level for the test of homogeneity of variance is the p-value for the ANOVA F-test on the dispersion variable. All of the homogeneity of variance tests available in PROC GLM except Bartlett's use this approach.

If one of these tests rejects the assumption of homogeneity of variance, you should use Welch's ANOVA (which is a very messy formula. I certainly would never expect anyone to memorize this; but $G = 2$, this test is equivalent to Welch's t-test: $\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$) instead of the usual ANOVA to test for differences between group means.

```
proc glm data=fev2;
class center;
model fev=center/noint;
means center/hovtest=bf welch;
run;
```

```
                          The GLM Procedure
Brown and Forsythes Test for Homogeneity of fev Variance
           ANOVA of Absolute Deviations from Group Medians
                          Sum of        Mean
    Source        DF     Squares      Square    F Value    Pr > F
    center         2     0.00818     0.00409       0.04    0.9586
    Error         57      5.5105      0.0967


                     Welchs ANOVA for fev
               Source          DF    F Value    Pr > F
               center      2.0000       3.02    0.0614
               Error      35.5447
```

## How do we do this in R?

```
> library(car)
> FEV = read.table("FEV2.dat", header = T)
> head(FEV)

   center  fev
1 hopkins 3.23
2 hopkins 3.47
3 hopkins 1.86
4 hopkins 2.47
5 hopkins 3.01
6 hopkins 1.69

> table(FEV$center)

 hopkins ranchola  stlouis
      21       16       23

> summary(FEV$fev)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.690   2.515   2.830   2.831   3.220   4.060

> # cell means predictor matrix
> X_cell = model.matrix(~ factor(center)-1, data = FEV)
```

```
> head(X_cell, 3)

  factor(center)hopkins factor(center)ranchola factor(center)stlouis
1                     1                      0                     0
2                     1                      0                     0
3                     1                      0                     0

> tail(X_cell, 3)

   factor(center)hopkins factor(center)ranchola factor(center)stlouis
58                     0                      0                     1
59                     0                      0                     1
60                     0                      0                     1

> # factor() tells R center is categorical (factor variable),
> # and -1 suppresses the intercept
```

## We can specify this cell means coding directly in lm()

```
> fit = lm(fev ~ factor(center)-1, data = FEV)
> summary(fit) # similar to SAS

Call:
lm(formula = fev ~ factor(center) - 1, data = FEV)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32250 -0.32250 -0.02244  0.32630  1.18130

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
factor(center)hopkins    2.6262     0.1100   23.88   <2e-16 ***
factor(center)ranchola   3.0325     0.1260   24.07   <2e-16 ***
factor(center)stlouis    2.8787     0.1051   27.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.504 on 57 degrees of freedom
Multiple R-squared:  0.9709,        Adjusted R-squared:  0.9693
F-statistic: 633.2 on 3 and 57 DF,  p-value: < 2.2e-16
```

## Now lets check for HOV

```
> leveneTest(y=FEV$fev, group=FEV$center, center=median)

Levenes Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.0423 0.9586
      57

> # if reject the null, Welchs ANOVA
> oneway.test(fev ~ center,
+             data=FEV,
+             var.equal=FALSE)

        One-way analysis of means (not assuming equal variances)

data:  fev and center
F = 3.0211, num df = 2.000, denom df = 35.545, p-value = 0.06141
```

What if we wanted to perform the overall test, where we test for the equality of means? Specify contrast matrix for cell means coding

```
> # overall test
> C = matrix(c(1, -1, 0, 1, 0, -1), nrow = 2, byrow = T)
> print(C)

     [,1] [,2] [,3]
[1,]    1   -1    0
[2,]    1    0   -1

> linearHypothesis(fit, C)

Linear hypothesis test

Hypothesis:
factor(center)hopkins - factor(center)ranchola = 0
factor(center)hopkins - factor(center)stlouis = 0

Model 1: restricted model
Model 2: fev ~ factor(center) - 1

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     59 16.063
```

```
2     57 14.480  2    1.5828 3.1153  0.052 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now peform pairwise comparison of means with Scheffe p-value correction

```
> library(DescTools)
> ScheffeTest(aov(fev ~ factor(center), data = FEV))

  Posthoc multiple comparisons of means : Scheffe Test
    95% family-wise confidence level

$factor(center)
                       diff      lwr.ci    upr.ci   pval
ranchola-hopkins  0.4063095 -0.01408874 0.8267078 0.0603 .
stlouis-hopkins   0.2525052 -0.12986364 0.6348740 0.2605
stlouis-ranchola -0.1538043 -0.56622280 0.2586141 0.6466

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

## What if we wanted to use reference cell coding?

```
> fit = lm(fev ~ factor(center), data = FEV)
> summary(fit)

Call:
lm(formula = fev ~ factor(center), data = FEV)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32250 -0.32250 -0.02244  0.32630  1.18130

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.6262     0.1100  23.877   <2e-16 ***
factor(center)ranchola   0.4063     0.1673   2.429   0.0183 *
factor(center)stlouis    0.2525     0.1521   1.660   0.1024
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.504 on 57 degrees of freedom
Multiple R-squared:  0.09854,        Adjusted R-squared:  0.06691
F-statistic: 3.115 on 2 and 57 DF,  p-value: 0.052
```

```
> # overall test for regression
> anova(fit)

Analysis of Variance Table

Response: fev
               Df  Sum Sq Mean Sq F value Pr(>F)
factor(center)  2  1.5828 0.79142  3.1153  0.052 .
Residuals      57 14.4803 0.25404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the overall corrected test for regression here is the same as the overall test in cell means encoding (utilizing the proper contrasts). The overall corrected test of regression in cell means coding is testing what exactly?

If you want to choose a different reference level, you can specify a different reference level in center using relevel()

```
> FEV$center = relevel(FEV$center, ref = "ranchola")
> fit = lm(fev ~ factor(center), data = FEV)
> summary(fit)

Call:
lm(formula = fev ~ factor(center), data = FEV)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32250 -0.32250 -0.02244  0.32630  1.18130

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.0325     0.1260  24.066   <2e-16 ***
factor(center)hopkins  -0.4063     0.1673  -2.429   0.0183 *
factor(center)stlouis  -0.1538     0.1641  -0.937   0.3525
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.504 on 57 degrees of freedom
```

```
Multiple R-squared:  0.09854,       Adjusted R-squared:  0.06691
F-statistic: 3.115 on 2 and 57 DF,  p-value: 0.052
```

## Next: Two-way ANOVA

- Muller and Fetterman, Chapter 14: "Complete, Two-Way Factorial ANOVA"

# Lecture 16: Two-Way ANOVA

*Reading Assignment:*

- Muller and Fetterman, Chapter 14: "Complete, Two-Way Factorial ANOVA" (Required)

In two-way analysis of variance (ANOVA), we wish to evaluate the importance of all combinations of two categorical variables in predicting a Gaussian response.

First, we define several terms commonly used to describe studies well-suited to ANOVA:

**Cell**: group of observational units, all receiving same treatment.

**Balanced**: all cells (present) have the same number of observational units.

**Complete**: all cells have at least one observation.

Balanced, complete designs are the simplest and lead to unambiguous tests. We use this situation as our prototype. Other designs require careful treatment and thoughtful consideration, though they can be handled in the general linear model framework.

## Model Concepts

### Factorial Design

A **factorial** design includes all combinations of the levels of two or more treatments.

If factor A has $a$ levels and factor B has $b$ levels, then a two-way factorial contains $a \cdot b$ combinations (cells).

For example, let factor A represent drug dose ($a = 2$) and let factor B represent drug formulations ($b = 3$) for a new cholesterol-reducing compound.

## Two-Way Factorial Design

$N$ Replicates

| | $DRUG_1$ | $DRUG_2$ | $DRUG_3$ |
|---|---|---|---|
| $DOSE_1$ | | | |
| $DOSE_2$ | | | |

Treatments in a factorial are completely crossed (with each other). This means that each drug is tested at each dose, and *vice versa*.

One observation per cell ($a \cdot b = 6$) gives one replicate. For a balanced design, let $N$ equal the # of observations in each cell so that $n = Nab$. So for the balanced two-way factorial design in the previous table, each dose-drug combination is tested in $N$ patients for a total of $6N$ patients in the study.

We generally like to see 5-10 replicates for any analysis (this allows estimation of interaction terms), though further assumptions may be made in order to make inferences for smaller studies.

## Cell and Marginal Means

Define a row marginal mean for balanced data as $\mu_{i\cdot} = \sum_{j=1}^{b} \mu_{ij}/b$, which is a row average in the table, where $\mu_{ij}$ is the mean in the cell in row $i$ and column $j$ of the table.

Define a column marginal mean for balanced data as $\mu_{\cdot j} = \sum_{i=1}^{a} \mu_{ij}/a$, which is a column average across the rows.

Note that the first subscript indicates the row (first factor) and the second subscript indicates the column (second factor).

The marginal means of factor A, $\{\mu_{1\cdot}, \mu_{2\cdot}\}$, have been averaged across factor B, and the marginal means of factor B, $\{\mu_{\cdot 1}, \mu_{\cdot 2}, \mu_{\cdot 3}\}$, have been averaged across factor A.

Consider the following marginal means for the cholesterol drug example.

Two-Way Factorial Design

| | $DRUG_1$ | $DRUG_2$ | $DRUG_3$ | |
|---|---|---|---|---|
| $DOSE_1$ | $\mu_{11}$ | $\mu_{12}$ | $\mu_{13}$ | $\mu_{1\cdot}$ |
| $DOSE_2$ | $\mu_{21}$ | $\mu_{22}$ | $\mu_{23}$ | $\mu_{2\cdot}$ |
| | $\mu_{\cdot 1}$ | $\mu_{\cdot 2}$ | $\mu_{\cdot 3}$ | $\mu_{\cdot\cdot}$ |

1. $\mu_{12}$ is the population mean cholesterol for the $N$ subjects taking dose 1 of drug 2,
2. $\mu_{1\cdot}$ is the population mean cholesterol for the $3N$ patients taking dose 1, averaged over all three drugs,
3. $\mu_{\cdot 2}$ is the population mean cholesterol for the $2N$ patients taking drug 2, averaged over the two doses, and
4. $\mu_{\cdot\cdot}$ is the population mean cholesterol over all $n = 6N$ subjects in the study.

## Primary Hypotheses

In a model without *interaction*, main effects describe marginal means. To test equality of marginal means, we have the hypotheses:

$$H_0 : \mu_{1\cdot} = \mu_{2\cdot} = \ldots = \mu_{a\cdot}$$

and

$$H_0 : \mu_{\cdot 1} = \mu_{\cdot 2} = \ldots = \mu_{\cdot b},$$

which correspond to testing if the factor A means are the same and if the factor B means are the same, respectively. These are called the tests of the *main effects* of factors A and B.

In the absence of interaction, we test main effects using contrasts like those in one-way ANOVA.

The *interaction* describes the difference among differences (is the dose effect the same for all three drugs?). If the interaction is significant, then the effect of factor B depends on the level of factor A or *vice versa*. Consider the hypothesis that the dose effect of drug A is the same as the dose effect of drug B:

$$H_0: \qquad \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22}$$

$$\Longleftrightarrow$$

$$H_0: \qquad (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) = 0.$$

If the null hypothesis is true, then the change in cholesterol going from dose 1 to dose 2 on drug A is the same as the change in cholesterol going from dose 1 to dose 2 on drug B. If the null hypothesis is false, then the dose effect is larger for one drug than for the other.

## Coding and Computation

How *do* we code an interaction model? If we use classical ANOVA coding (like SAS), we have the model

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

$i = 1, 2$, $j = 1, 2, 3$, $k = 1, \ldots, N$. In this case, we have $2 \times 3 = 6$ dose-drug combinations but $1 + 2 + 3 + 6 = 12$ parameters! This model is grossly overparameterized, since we can fit only 6 cell means (thus 6 primary parameters are not estimable).

We could use cell means and fit the model

$$E(y_{ijk}) = \gamma_{ij},$$

in which each $\gamma$ represents a cell mean for a given dose-drug combination. A more popular approach is to use reference cell coding, which sets one level of each drug and dose to be the reference. For the above example, if we set the highest level of each variable to be the

reference, we have the model

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

$i = 1$, $j = 1, 2$, $k = 1, \ldots, N$. With only these six parameters in the model, everything is now estimable.

## Reference Cell (Regression) Coding

Consider reference cell coding for our cholesterol drug example. Suppose that we have two replicates and that our reference dose is dose 2 and our reference drug is drug 3. There is one column of 1's for the reference cell mean, $a - 1$ columns for the main effect of factor A, $b - 1$ columns for the main effect of factor B, and $(a - 1)(b - 1)$ columns for the interaction.

$$
\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \end{bmatrix}
=
\begin{bmatrix}
1 & \vdots & 1 & \vdots & 1 & 0 & \vdots & 1 & 0 \\
1 & \vdots & 1 & \vdots & 1 & 0 & \vdots & 1 & 0 \\
1 & \vdots & 1 & \vdots & 0 & 1 & \vdots & 0 & 1 \\
1 & \vdots & 1 & \vdots & 0 & 1 & \vdots & 0 & 1 \\
1 & \vdots & 1 & \vdots & 0 & 0 & \vdots & 0 & 0 \\
1 & \vdots & 1 & \vdots & 0 & 0 & \vdots & 0 & 0 \\
1 & \vdots & 0 & \vdots & 1 & 0 & \vdots & 0 & 0 \\
1 & \vdots & 0 & \vdots & 1 & 0 & \vdots & 0 & 0 \\
1 & \vdots & 0 & \vdots & 0 & 1 & \vdots & 0 & 0 \\
1 & \vdots & 0 & \vdots & 0 & 1 & \vdots & 0 & 0 \\
1 & \vdots & 0 & \vdots & 0 & 0 & \vdots & 0 & 0 \\
1 & \vdots & 0 & \vdots & 0 & 0 & \vdots & 0 & 0
\end{bmatrix}
\begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix}
+ \boldsymbol{\varepsilon}
$$

The interpretations of the parameters may be obtained by considering $E(y_{ij})$ for subjects on various doses (indexed by $i$) of various drugs (indexed by $j$). This means, for example, that we no longer have one parameter in our model that describes the dose effect – instead, we have 3 parameters that describe the dose effect; one for each drug.

In this example, $\mu$ represents the average cholesterol for subjects at dose 2 of drug 3. $\alpha_1$ is the difference in cholesterol for subjects on dose 1 of drug 3 compared to subjects on dose 2 of drug 3. $\beta_1$ is the difference in cholesterol for subjects on dose 2 of drug 1 compared to subjects on dose 2 of drug 3. $\beta_2$ is the difference in cholesterol for subjects on dose 2 of drug 2 compared to subjects on dose 2 of drug 3.

The difference in dose effects for drug 1 versus drug 3 is given by $\gamma_{11}$, and the difference in dose effects for drug 2 versus drug 3 is given by $\gamma_{12}$.

## Generating Cell Means

$\boldsymbol{C} =$Es$(\boldsymbol{X})$ gives the cell means, $\boldsymbol{C\beta} =$Es$(\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\theta} = \{\mu_{jk}\}$, for complete designs, whether balanced or not.

For the cholesterol example, the cell means are provided below.

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix},$$

so that we have the following cell means.

| *Dose* | *Drug* | *Mean* |
|---|---|---|
| 1 | 1 | $\mu + \alpha_1 + \beta_1 + \gamma_{11}$ |
| 1 | 2 | $\mu + \alpha_1 + \beta_2 + \gamma_{12}$ |
| 1 | 3 | $\mu + \alpha_1$ |
| 2 | 1 | $\mu + \beta_1$ |
| 2 | 2 | $\mu + \beta_2$ |
| 2 | 3 | $\mu$ |

## Computing Estimates and Tests

### Balanced, Complete Designs with Equal Cell Size

We can use simple formulae for hand computation in balanced designs, leading to simple interpretations.

| Source | df | SS |
|---|---|---|
| A | $a-1$ | $bN\sum_{i=1}^{a}(\overline{y}_{i..}-\overline{y}_{...})^2$ |
| B | $b-1$ | $aN\sum_{j=1}^{b}(\overline{y}_{.j.}-\overline{y}_{...})^2$ |
| AB | $(a-1)(b-1)$ | $N\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...})^2$ |
| Error | $ab(N-1)$ | $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{N}(y_{ijk}-\overline{y}_{ij.})^2$ |
| Total | $abN-1$ | $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{N}(y_{ijk}-\overline{y}_{...})^2$ |

The $SS$ formulae presented here are correct only with complete, balanced data, unless explicitly stated otherwise.

### Unbalanced, Complete Data

With unbalanced, complete data, we revert to the general linear model framework and fit $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. By specifying the appropriate contrast matrices, we are able to carry out the appropriate tests.

One must fit multiple models in order to generate added-in-order tests.

## Contrast Matrices for Marginal Means

### Mean Estimation with Reference Cell Coding

Start with the matrix which computes cell means that correspond to the interaction in our cholesterol example: $\boldsymbol{C}_{\mathsf{A}\times\mathsf{B}} = \mathsf{Es}(\boldsymbol{X})$

$\boldsymbol{C}_{\mathsf{A}\times\mathsf{B}}\boldsymbol{\beta} = \mathsf{Es}(\boldsymbol{X})\,\boldsymbol{\beta} = \{\mu_{jk}\}$

$$\boldsymbol{\theta}_{\mathsf{A}\times\mathsf{B}} = \{\mu_{jk}\} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{23} \end{bmatrix}.$$

To estimate the marginal means of factor A (dose) for our example, we use

$$\boldsymbol{C}_{\text{A ref}} = \begin{bmatrix} 1 & \vdots & 1 & \vdots & \frac{1}{3} & \frac{1}{3} & \vdots & \frac{1}{3} & \frac{1}{3} \\ & & & & & & & & \\ 1 & \vdots & 0 & \vdots & \frac{1}{3} & \frac{1}{3} & \vdots & 0 & 0 \end{bmatrix}$$

to obtain $\boldsymbol{\theta}_A = \begin{bmatrix} \mu_{1\cdot} \\ \mu_{2\cdot} \end{bmatrix}$.

This becomes clear if we think about averaging the expected values by dose.

Similarly, for marginal means of factor B (drug), we have

$$\boldsymbol{C}_{\text{B ref}} = \begin{bmatrix} 1 & \vdots & \frac{1}{2} & \vdots & 1 & 0 & \vdots & \frac{1}{2} & 0 \\ \\ 1 & \vdots & \frac{1}{2} & \vdots & 0 & 1 & \vdots & 0 & \frac{1}{2} \\ \\ 1 & \vdots & \frac{1}{2} & \vdots & 0 & 0 & \vdots & 0 & 0 \end{bmatrix}, \text{ leading to}$$

$$\boldsymbol{\theta}_{\text{B}} = \begin{bmatrix} \mu_{\cdot 1} \\ \mu_{\cdot 2} \\ \mu_{\cdot 3} \end{bmatrix}.$$

The grand mean can be thought of as the average of all cell means or as the average of either set of marginal means:

$$\boldsymbol{C}_{\text{grand mean ref}} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \end{bmatrix}.$$

## Testing with Reference Cell Coding

One must take care when testing the main effects of factor A or factor B in the presence of their interaction. Specifically, testing that the $\alpha$'s are equal to zero (or that the $\beta$'s are equal to zero) is not testing marginal means. In fact, this would be a test of simple main effects *at the reference level* of the other factor.

For factor A, we really wish to test the hypothesis

$$H_0 : \mu_{1\cdot} = \mu_{2\cdot}$$

Thus we wish to test the hypothesis

$$H_0 : \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_{1\cdot} \\ \mu_{2\cdot} \end{bmatrix} = 0.$$

Thus we may define the appropriate contrast matrix $\mathbf{C}$ for factor A as

$$
\begin{aligned}
\mathbf{C} &= \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & \vdots & 1 & \vdots & \frac{1}{3} & \frac{1}{3} & \vdots & \frac{1}{3} & \frac{1}{3} \\ & & & & & & & & \\ 1 & \vdots & 0 & \vdots & \frac{1}{3} & \frac{1}{3} & \vdots & 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \vdots & 1 & \vdots & 0 & 0 & \vdots & \frac{1}{3} & \frac{1}{3} \end{bmatrix}.
\end{aligned}
$$

Similarly, for testing marginal means of factor B, we have the contrast matrix

$$
\begin{aligned}
\mathbf{C} &= \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & \vdots & \frac{1}{2} & \vdots & 1 & 0 & \vdots & \frac{1}{2} & 0 \\ 1 & \vdots & \frac{1}{2} & \vdots & 0 & 1 & \vdots & 0 & \frac{1}{2} \\ 1 & \vdots & \frac{1}{2} & \vdots & 0 & 0 & \vdots & 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & \vdots & 0 & \vdots & 1 & 0 & \vdots & \frac{1}{2} & 0 \\ 0 & \vdots & 0 & \vdots & 0 & 1 & \vdots & 0 & \frac{1}{2} \end{bmatrix}.
\end{aligned}
$$

The test of interaction is simply a test that the interaction terms $\boldsymbol{\gamma}$ are equal to zero.

## Choosing and Interpreting Tests

### Model Comparisons

All GLH tests correspond to comparing two models. Note that marginal means tests in reference cell coding do not correspond to column deletions.

### Testing Strategies

All testing strategies are special cases of those in Chapters 5 and 11. The recommended strategy is a backwards group-wise approach. The first group contains the main effect of A variables, the second group contains the main effect of B variables, and the third group contains the interaction variables. Start by testing the interaction. If significant, stop, and fit the full model. If not significant, drop the interaction. If the interaction is dropped, proceed with testing of the main effects.

## Step-Down Tests

### Simple Main Effect Tests

A simple main effect equals a difference in cell means within a column ($\mu_{jk} - \mu_{j'k}$) or within a row ($\mu_{jk} - \mu_{jk'}$). These tests are of interest if the interaction terms are significant. The simple main effect (SME) test evaluates one effect while holding the other constant. There exist $b$ SME tests for factor A and $a$ SME tests for factor B. In our example, we may wish to use a simple main effect test to determine, for example, if there is a dose effect for drug A.

### Second Level Step-Down Tests

If a SME with two or more numerator degrees of freedom is significant, additional tests are required to determine where the difference lies.

## Missing Data

The pattern of cell sizes may have profound effects on computation and interpretation of test statistics in ANOVA. The complete, balanced case involves no missing data. The complete, unbalanced case involves at least some missing data, meaning that additional observations would be needed to make the design balanced. Careful specification of $\boldsymbol{C}$ for marginal means tests "automatically" provides valid solutions with complete unbalanced data. The model corresponds to simply deleting some rows in $\boldsymbol{X}$ and $\boldsymbol{y}$ when missing data is simply due to some lack of balance. In this case, we have a model like

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{J}_{n_1} & & \\ & \boldsymbol{J}_{n_2} & \\ & & \boldsymbol{J}_{n_3} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \boldsymbol{\varepsilon}$$

with $n = n_1 + n_2 + n_3$. For this model $\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\mu}_1 & \widehat{\mu}_2 & \widehat{\mu}_3 \end{bmatrix}'$.

Strongly unbalanced data create additional difficulties.

With unbalanced data avoid automatic coding schemes provided by computer programs (such as the CLASS statement in PROC GLM).

## More About Missing Data

The above comments apply when data is unbalanced *by design*. Suppose we originally planned to treat 50 patients on dose 1 and 40 patients on dose 2. Alternatively, suppose that originally 50 patients were taking each dose, but 10 patients taking dose 2 dropped out of the study due to toxic side effects at that dose. In the second case, a *complete case analysis*, or an analysis that deletes some rows in $\mathbf{X}$ and $\mathbf{y}$, may lead to biased parameter estimates.

The validity of any analysis with incomplete data depends on the reasons for which data (response or covariates) are missing.

A *complete case analysis* is valid only when data are *missing completely at random (MCAR)*, which means that the completely observed subjects are a completely random sample of all the subjects in the study. Complete case analysis is the default analysis in SAS PROC REG and SAS PROC GLM when data (responses or covariates) are missing, regardless of its validity for the data in question.

Other types of missingness include *missing at random (MAR)*, which means that missingness does not depend on the values of the variables that are missing but may depend on observed quantities. Examples of MAR mechanisms

- A subject may be removed from a trial if his/her condition is not controlled sufficiently well (according to pre-defined criteria on the

response).

- Two measurements of the same variable are made at the same time. If they differ by more than a given amount a third is taken. This third measurement is missing for those that do not differ by the given amount.

*Nonignorable missingness*, which means that the missingness may depend on the values of the variables that are missing.

In the latter two cases, the *complete case analysis* is not valid in general and may lead to considerable bias and faulty results. If a *complete case analysis* is carried out, deleted observations must be reported, and the investigator must seriously consider the validity of the analysis.

Even in the "ideal" MCAR situation, replacing missing data with simple imputed values (such as the average for all subjects) biases the

estimate of error variance downward and therefore inflates the type I error rate. Despite this relatively well-known fact, many investigators substitute means for missing values by default. "Making up" data in this manner can have extremely severe consequences and should be avoided.

Recently, SAS has implemented PROC MI and PROC MIANALYZE, which carry out multiple imputation, a more sophisticated method for handling missing observations. However, these procedures are still in the development stages and are not yet available for a wide class of problems (currently, PROC MI requires all variables to follow Gaussian distributions for general missing data patterns).

## Example: Automobile Pollution Filter Noise

Associated Octel Company developed an automobile silencing filter, the Octel filter, that Texaco, Inc. stated (before the Air and Water Pollution Subcommittee of the Senate Public Works Committee) was just as effective as the standard silencing filter. The president of Texaco at the time presented data that we now analyze to back up his claim that the new filter (TYPE1=1 for standard, TYPE1=0 for Octel) was no more noisy than the standard filter. The dependent variable, NOISE, is noise level in decibels, with factor A as the vehicle size (small, medium, or large, with large as the reference level) and factor B as filter type (standard or Octel, with Octel as the reference level).

Our goal is to evaluate the noise levels of the three car sizes at the two filter types.

First, we define the following variables for the reference cell coding scheme.

- SIZE1, which takes the value 1 for small cars and 0 otherwise
- SIZE2, which takes the value 1 for medium cars and 0 otherwise
- S1TYPE1, which takes the value 1 for small cars with the standard filter and zero otherwise
- S2TYPE1, which takes the value 1 for medium cars with the standard filter and zero otherwise

Next, let's take a look at the mean noise levels for each car size and filter type.

| Size | Silencer/Filter Type | Average Noise Level |
|---|---|---|
| Small | Standard | 825.83 |
| Small | Octel | 822.50 |
| Medium | Standard | 845.83 |
| Medium | Octel | 821.67 |
| Large | Standard | 775.00 |
| Large | Octel | 770.00 |

What is your impression based on these cell means?

Next, we fit the model

$$\begin{aligned} NOISE &= \beta_0 + \beta_1 SIZE1 + \beta_2 SIZE2 + \beta_3 TYPE1 + \\ &\quad \beta_4 S1TYPE1 + \beta_5 S2TYPE1 + \varepsilon. \end{aligned}$$

The SAS code is below.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
estimate Grand Mean
        intercept 6 size1 2 size2 2 type1 3 s1type1 1 s2type1 1/divisor=6;
estimate Marg Mean: Small
        intercept 2 size1 2 size2 0 type1 1 s1type1 1 s2type1 0/divisor=2;
estimate Marg Mean: Medium
        intercept 2 size1 0 size2 2 type1 1 s1type1 0 s2type1 1/divisor=2;
estimate Marg Mean: Large
        intercept 2 size1 0 size2 0 type1 1 s1type1 0 s2type1 0/divisor=2;
estimate Marg Mean: Standard
        intercept 3 size1 1 size2 1 type1 3 s1type1 1 s2type1 1/divisor=3;
estimate Marg Mean: Octel
        intercept 3 size1 1 size2 1 type1 0 s1type1 0 s2type1 0/divisor=3;

contrast Interaction Silencer by Size
        intercept 0 size1 0 size2 0 type1 0 s1type1 1 s2type1 0,
        intercept 0 size1 0 size2 0 type1 0 s1type1 0 s2type1 1;
```

```
contrast Main Effect Vehicle Size
        intercept 0 size1 2 size2 0 type1 0 s1type1 1 s2type1 0,
        intercept 0 size1 0 size2 2 type1 0 s1type1 0 s2type1 1;

contrast Main Effect Silencer
        intercept 0 size1 0 size2 0 type1 3 s1type1 1 s2type1 1;
run;
```

```
/* ALTERNATIVELY, you can use the code below */
/* DANGER DANGER -- this code uses the class statement */
proc glm;
class size type;
model noise=size type size*type;
run;
```

The output is provided below.

```
                          The GLM Procedure

Dependent Variable: noise

                                        Sum of
Source                     DF          Squares     Mean Square    F Value
Model                       5      27911.80556      5582.36111      85.34
Error                      30       1962.50000        65.41667

Corrected Total            35      29874.30556

                    Source                  Pr > F
                    Model                   <.0001
                    Error

                    Corrected Total
```

| R-Square | Coeff Var | Root MSE | noise Mean |
|---|---|---|---|
| 0.934308 | 0.998354 | 8.088057 | 810.1389 |

| Source | DF | Type I SS | Mean Square | F Value |
|---|---|---|---|---|
| size1 | 1 | 3542.01389 | 3542.01389 | 54.15 |
| size2 | 1 | 22509.37500 | 22509.37500 | 344.09 |
| type1 | 1 | 1056.25000 | 1056.25000 | 16.15 |
| s1type1 | 1 | 253.12500 | 253.12500 | 3.87 |
| s2type1 | 1 | 551.04167 | 551.04167 | 8.42 |

| Source | Pr > F |
|---|---|
| size1 | <.0001 |
| size2 | <.0001 |
| type1 | 0.0004 |
| s1type1 | 0.0585 |
| s2type1 | 0.0069 |

| Source | DF | Type III SS | Mean Square | F Value |
|---|---|---|---|---|
| size1 | 1 | 8268.750000 | 8268.750000 | 126.40 |
| size2 | 1 | 8008.333333 | 8008.333333 | 122.42 |
| type1 | 1 | 75.000000 | 75.000000 | 1.15 |

| Source | Pr > F |
|---|---|
| size1 | <.0001 |
| size2 | <.0001 |
| type1 | 0.2928 |

| Source | DF | Type III SS | Mean Square | F Value |
|---|---|---|---|---|
| s1type1 | 1 | 4.166667 | 4.166667 | 0.06 |
| s2type1 | 1 | 551.041667 | 551.041667 | 8.42 |

| Source | Pr > F |
|---|---|
| s1type1 | 0.8025 |
| s2type1 | 0.0069 |

| Contrast | DF | Contrast SS |
|---|---|---|
| Interaction Silencer by Size | 2 | 804.16667 |
| Main Effect Vehicle Size | 2 | 26051.38889 |
| Main Effect Silencer | 1 | 1056.25000 |

| Contrast | Mean Square | F Value |
|---|---|---|
| Interaction Silencer by Size | 402.08333 | 6.15 |
| Main Effect Vehicle Size | 13025.69444 | 199.12 |
| Main Effect Silencer | 1056.25000 | 16.15 |

| Contrast | Pr > F |
|---|---|
| Interaction Silencer by Size | 0.0058 |
| Main Effect Vehicle Size | <.0001 |
| Main Effect Silencer | 0.0004 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Grand Mean | 810.138889 | 1.34800951 | 600.99 | <.0001 |
| Marg Mean: Small | 824.166667 | 2.33482095 | 352.99 | <.0001 |
| Marg Mean: Medium | 833.750000 | 2.33482095 | 357.09 | <.0001 |
| Marg Mean: Large | 772.500000 | 2.33482095 | 330.86 | <.0001 |
| Marg Mean: Standard | 815.555556 | 1.90637333 | 427.80 | <.0001 |

| | | | | |
|---|---|---|---|---|
| Marg Mean: Octel | 804.722222 | 1.90637333 | 422.12 | <.0001 |

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 770.0000000 | 3.30193546 | 233.20 | <.0001 |
| size1 | 52.5000000 | 4.66964191 | 11.24 | <.0001 |
| size2 | 51.6666667 | 4.66964191 | 11.06 | <.0001 |
| type1 | 5.0000000 | 4.66964191 | 1.07 | 0.2928 |
| s1type1 | -1.6666667 | 6.60387092 | -0.25 | 0.8025 |
| s2type1 | 19.1666667 | 6.60387092 | 2.90 | 0.0069 |

Using the SAS output, we can construct the ANOVA table as follows.

| Source | df | SS | MS | F | p |
|---|---|---|---|---|---|
| Size | | | | | |
| Silencer | | | | | |
| Size*Silencer | | | | | |
| Error | | | | | |
| Total | | | | | |

What further tests, if any, should we consider?

Because the interaction is significant, a logical next step is to conduct simple main effect tests of the filter effect within each car size. The SAS code below may be used to conduct these SME tests.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
contrast SME Silencer at Size Small
        intercept 0 size1 0 size2 0 type1 1 s1type1 1 s2type1 0;
contrast SME Silencer at Size Medium
        intercept 0 size1 0 size2 0 type1 1 s1type1 0 s2type1 1;
contrast SME Silencer at Size Large
        intercept 0 size1 0 size2 0 type1 1 s1type1 0 s2type1 0;
run;
```

The additional output is provided below.

| Contrast | DF | Contrast SS |
|---|---|---|
| SME Silencer at Size Small | 1 | 33.33333 |
| SME Silencer at Size Medium | 1 | 1752.08333 |
| SME Silencer at Size Large | 1 | 75.00000 |

| Contrast | Mean Square | F Value |
|---|---|---|

| | | |
|---|---|---|
| SME Silencer at Size Small | 33.33333 | 0.51 |
| SME Silencer at Size Medium | 1752.08333 | 26.78 |
| SME Silencer at Size Large | 75.00000 | 1.15 |

| Contrast | Pr > F |
|---|---|
| SME Silencer at Size Small | 0.4808 |
| SME Silencer at Size Medium | <.0001 |
| SME Silencer at Size Large | 0.2928 |

What do you conclude?

Alternatively (less logically?), you could have tested the SME of car size for each filter type. The SAS code and additional output are provided below.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
contrast SME Size at Standard Silencer
        intercept 0 size1 1 size2 -1 type1 0 s1type1 1 s2type1 -1,
        intercept 0 size1 1 size2 0 type1 0 s1type1 1 s2type1 0;
contrast SME Size at Octel Silencer
        intercept 0 size1 1 size2 -1 type1 0 s1type1 0 s2type1 0,
        intercept 0 size1 1 size2 0 type1 0 s1type1 0 s2type1 0;
run;
*************************************************************
```

| Contrast | DF | Contrast SS |
|---|---|---|
| SME Size at Standard Silencer | 2 | 16002.77778 |
| SME Size at Octel Silencer | 2 | 10852.77778 |

| Contrast | Mean Square | F Value |
|---|---|---|

| | | |
|---|---|---|
| SME Size at Standard Silencer | 8001.38889 | 122.31 |
| SME Size at Octel Silencer | 5426.38889 | 82.95 |

| Contrast | Pr > F |
|---|---|
| SME Size at Standard Silencer | <.0001 |
| SME Size at Octel Silencer | <.0001 |

What do you conclude?

Because the SME tests for size were significant for both filter types, we can now carry out step-down tests to determine specifically which car type means differ for each filter type.

```
proc glm;
model noise=size1 size2 type1 s1type1 s2type1/solution;
contrast Step-Down of SME at Std: Small vs. Med
        intercept 0 size1 1 size2 -1 type1 0 s1type1 1 s2type1 -1;
contrast Step-Down of SME at Std: Small vs. Large
        intercept 0 size1 1 size2 0 type1 0 s1type1 0 s2type1 0;
contrast Step-Down of SME at Std: Med vs. Large
        intercept 0 size1 0 size2 1 type1 0 s1type1 0 s2type1 1;
contrast Step-Down of SME at Octel: Small vs. Med
        intercept 0 size1 1 size2 -1 type1 0 s1type1 0 s2type1 0;
contrast Step-Down of SME at Octel: Small vs. Large
        intercept 0 size1 1 size2 0 type1 0 s1type1 0 s2type1 0;
contrast Step-Down of SME at Octel: Med vs. Large
        intercept 0 size1 0 size2 1 type1 0 s1type1 0 s2type1 0;
run;
```

**********

| Contrast | DF | Contrast SS |
|---|---|---|
| Step-Down of SME at Std: Small vs. Med | 1 | 1200.00000 |
| Step-Down of SME at Std: Small vs. Large | 1 | 8268.75000 |
| Step-Down of SME at Std: Med vs. Large | 1 | 15052.08333 |

| Contrast | Mean Square | F Value |
|---|---|---|
| Step-Down of SME at Std: Small vs. Med | 1200.00000 | 18.34 |
| Step-Down of SME at Std: Small vs. Large | 8268.75000 | 126.40 |
| Step-Down of SME at Std: Med vs. Large | 15052.08333 | 230.10 |

| Contrast | Pr > F |
|---|---|
| Step-Down of SME at Std: Small vs. Med | 0.0002 |
| Step-Down of SME at Std: Small vs. Large | <.0001 |
| Step-Down of SME at Std: Med vs. Large | <.0001 |

| Contrast | DF | Contrast SS |
|---|---|---|
| Step-Down of SME at Octel: Small vs. Med | 1 | 2.08333 |
| Step-Down of SME at Octel: Small vs. Large | 1 | 8268.75000 |
| Step-Down of SME at Octel: Med vs. Large | 1 | 8008.33333 |

| Contrast | Mean Square | F Value |
|---|---|---|
| Step-Down of SME at Octel: Small vs. Med | 2.08333 | 0.03 |
| Step-Down of SME at Octel: Small vs. Large | 8268.75000 | 126.40 |
| Step-Down of SME at Octel: Med vs. Large | 8008.33333 | 122.42 |

| Contrast | Pr > F |
|---|---|
| Step-Down of SME at Octel: Small vs. Med | 0.8596 |
| Step-Down of SME at Octel: Small vs. Large | <.0001 |
| Step-Down of SME at Octel: Med vs. Large | <.0001 |

What do you conclude?

For this analysis, we considered engine size as a *factor* instead of as a continuous or ordinal variable. Suppose instead we use one variable,

$$SIZE = \begin{cases} 1 & \text{small} \\ 2 & \text{medium} \\ 3 & \text{large} \end{cases}$$

to describe engine size. What are the implications for our analysis?

## Next: Logistic Regression

*Reading Assignment:*

- Weisberg Chapter 12 and KKMN, Chapter 23: "Logistic Regression Analysis"

# Lecture 17: Logistic Regression

*Reading Assignment:*

- Weisberg Chapter 12

Often, the response of interest in a scientific study is a binary variable, such as DISEASED/NOT DISEASED or DEAD/ALIVE. In this case, the linear regression model no longer holds because the errors will not follow a Gaussian distribution.

When studying linear regression, our models were of the form

$$E[\mathbf{y}] = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_{p-1} \mathbf{x}_{p-1}.$$

The response $\mathbf{y}$ was continuous, not discrete, and we wanted to predict the mean response and explain the variability among the observed outcomes.

When $\mathbf{y}$ is dichotomous, we observe only two possible values: "success" or "failure." Often, we use a 0/1 coding scheme so that "success" means $y = 1$ and "failure" means $y = 0$. We can talk about the mean of $\mathbf{y}$, which is the percentage of times that $y$ takes the value $1$, or the percentage of successes. We often denote this percentage as $p = E(y) = \text{Pr}(y = 1) = \text{Pr}(\text{success})$.

In this setting, we wish to estimate this probability $p$ as well as the effect of various explanatory variables or covariates on this success probability. In order to do so, we use the *logistic regression* model.

## Logistic Regression Model: Heuristics

Consider a study of cold incidence among French skiers (Pauling, PNAS, 1971), some of whom were given vitamin C.

| | COLD | NO COLD | Total |
|---|---|---|---|
| VIT C | 17 | 122 | 139 |
| NO VIT C | 31 | 109 | 140 |
| Total | 48 | 231 | 279 |

The *odds ratio* is a widely-used epidemiologic measure of association. It compares two or more groups in predicting the outcome variable. The *odds* are defined as the ratio of probabilities that an event (developing a cold) will occur divided by the probability that the same event will not occur. So if the probability of a cold is 0.25, then the odds of a cold are $\frac{0.25}{1.00-0.25} = \frac{1}{3}$. An odds of $\frac{1}{3}$ means that the

probability of a cold is one-third of the probability of no cold (in betting you often hear that the odds are "3 to 1" that the event will not occur). An *odds ratio* is just the ratio of two odds. When the *odds ratio* is one, then the two groups are equally likely to develop a cold.

In this sample, the probability of a cold for skiers taking vitamin C is $p_1 = \frac{17}{139} = 0.12$, and the corresponding probability for skiers not taking vitamin C is $p_2 = \frac{31}{140} = 0.22$. The *odds* of a cold for a skier taking vitamin C are $\frac{p_1}{1-p_1} = \frac{0.12}{1.00-0.12} = 0.14$, and for a skier not taking vitamin C, the odds are $\frac{p_2}{1-p_2} = \frac{0.22}{1.00-0.22} = 0.28$. We relate these odds using an *odds ratio*, defined as

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{0.14}{0.28} = 0.49$$

(after fixing some rounding error). This means that those skiers taking vitamin C have half the odds of a cold as skiers not taking vitamin C, or that vitamin C is protective against getting colds.

Aside: Why do we bother with the odds?

A **prospective** study watches for outcomes, such as the development of a disease, during the study period and relates this to other factors such as suspected risk or protection factor(s). The study usually involves taking a cohort of subjects and watching them over a long period. The outcome of interest should be common; otherwise, the number of outcomes observed will be too small to be statistically meaningful. All efforts should be made to avoid sources of bias such as the loss of individuals to follow up during the study. Prospective studies usually have fewer potential sources of bias and confounding than retrospective studies.

A **retrospective** study looks backwards and examines exposures to suspected risk or protection factors in relation to an outcome that is established at the start of the study. Many valuable case-control studies, such as Lane and Claypon's 1926 investigation of risk factors for breast cancer, were retrospective investigations. If the outcome of interest is uncommon, however, the size of prospective investigation

required to estimate relative risk is often too large to be feasible. In retrospective studies the odds ratio provides an estimate of relative risk. You should take special care to avoid sources of bias and confounding in retrospective studies.

A related measure, the *risk ratio*, is defined as

$$RR = \frac{p_1}{p_2}.$$

In this case,

$$RR = \frac{\frac{17}{139}}{\frac{31}{140}} = 0.55,$$

which is interpreted as "Vitamin C users have roughly half the risk of developing a cold." This measure of association has a simpler interpretation, but it usually cannot be estimated retrospectively because $p_1$ and $p_2$ cannot be estimated.

Specifically, with retrospective data, we cannot estimate

$p_1 = p(\text{disease} \mid \text{exposed})$ or $p_2 = p(\text{disease} \mid \text{unexposed})$. To estimate these quantities, we typically select a group of people based on exposure status and follow them through time to see whether or not they develop disease. In a case-control study, we select patients based on disease status and then determine exposure, so that we estimate $\pi_1 = p(\text{exposed} \mid \text{disease})$ and $\pi_2 = p(\text{exposed} \mid \text{no disease})$.

So how can we estimate the odds ratio, which is a function of $p_1$ and $p_2$, in case-control studies?

We wish to develop a statistical model for the cold data so that we can later incorporate other confounders, which might include family size or use of herbal remedies like echinacea.

A first strategy might be to fit a model using the form

$$E(y) = p = \beta_0 + \beta_1 x,$$

where $p$ is the probability of a cold, and $x$ is the vitamin C status, where $x = 1$ for vitamin C takers and 0 otherwise. This looks like an ordinary least squares regression model, in which the response (a probability) is continuous. However, this model is problematic because $p$ is restrained to lie in the interval $[0, 1]$, while $\beta_0 + \beta_1 x$ could technically take any value.

To ensure our estimate of $p$ is positive, we could try fitting the model

$$E(y) = p = \exp(\beta_0 + \beta_1 x),$$

but this model is also unsatisfactory because we could estimate $p$ to be greater than 1.

To solve both problems, we fit a model of the form

$$E(y) = p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x)}.$$

This *logistic function* cannot yield estimates of $p$ that are less than 0 or greater than 1.

We are not quite finished, though. This model is a little difficult to interpret; how does taking vitamin C affect your probability of getting a cold in this model? The relationship between $p$ and $x$ is clearly not linear! To make interpretation easier, we will use algebra to rewrite the

model.

$$
\begin{aligned}
p &= \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \iff \\
\frac{p}{1-p} &= \frac{\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}}{\frac{1}{1+\exp(\beta_0+\beta_1 x)}} \iff \\
\frac{p}{1-p} &= \exp(\beta_0 + \beta_1 x) \iff \\
\mathsf{log}\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 x \iff \\
\mathsf{logit}\,(p) &= \beta_0 + \beta_1 x.
\end{aligned}
$$

Thus $x$ is linearly related to the log-odds of a cold. The parameter $\beta_1$ is interpreted as the log odds ratio of a cold, and $\frac{\exp(\beta_0+\beta_1(1))}{\exp(\beta_0+\beta_1(0))} = \exp(\beta_1)$ provides the odds ratio.

For a continuous predictor $x$, the odds ratio between levels $i$ and $j$ of the predictor is given by $\frac{\exp(\beta_0+\beta_1(i))}{\exp(\beta_0+\beta_1(j))} = \exp(\beta_1(i-j))$. Thus the effect of increasing $x$ by the amount $d$ is to increase the odds that $y = 1$ by a factor of $\exp(\beta_1 d)$ or to increase the log odds that $y = 1$ by an increment of $\beta_1 d$.

## Logistic Regression Likelihood

Consider a study of the relationship between folic acid intake

$$x_1 = \begin{cases} 1 & \text{adequate folic acid intake} \\ 0 & \text{otherwise} \end{cases}$$

and preterm delivery

$$y = \begin{cases} 1 & \text{baby is born before 37 weeks} \\ 0 & \text{otherwise.} \end{cases}$$

Because race

$$x_2 = \begin{cases} 1 & \text{African American} \\ 0 & \text{otherwise} \end{cases}$$

is also related to the probability of preterm delivery and may be related to folic acid intake, we wish to control for it in our model.

The main question of interest is whether adequate folic acid intake is

associated with a reduced probability of preterm delivery. A secondary question may be how race is related to the probability of preterm delivery. (In the PIN study at UNC, African American women are at much lower risk of preterm delivery than are African American women in the general US population for reasons largely undetermined.)

The general logistic regression model is given by

$$\begin{aligned} \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} \end{aligned}$$

with $y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \ldots, n$, and the $y$'s independent of each other.

The likelihood is obtained as a product of the marginal distributions of

the $y$'s:

$$\begin{aligned} L(\mathbf{y} \mid p) &= \prod_{i=1}^{n} \mathsf{Pr}(y_i \mid p_i) \\ &= \prod_{i=1}^{n} \left[ p_i^{y_i} (1-p_i)^{1-y_i} \right], \end{aligned}$$

where $p_i$ is a function of the covariates $x_{i1}, \ldots, x_{i,p-1}$ and parameters $\beta_0, \ldots, \beta_{p-1}$ given by

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1})}.$$

Because this likelihood is a complex function of the elements of $\boldsymbol{\beta}$, maximizing it to find the maximum likelihood estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ requires an iterative procedure like Newton-Raphson or iteratively reweighted least squares. Computer programs (like SAS PROC LOGISTIC or SAS PROC GLM) provide maximum likelihood estimates of $\boldsymbol{\beta}$ as well as estimated large-sample covariances to be used for statistical inference.

## Test Statistics and Model Comparisons

For the logistic regression model, a test of

$$H_0 : \beta_k = 0$$

is a test of whether the $k^{th}$ covariate affects the probability of success, with the null hypothesis of probability of success independent of $x_k$.

To test this hypothesis for each variable in the model, SAS reports the Wald statistic, given by

$$Z = \frac{\widehat{\beta}_k}{\left(\widehat{\mathsf{Var}}(\widehat{\boldsymbol{\beta}})_{k+1,k+1}\right)^{\frac{1}{2}}},$$

which is asymptotically $N(0,1)$ under $H_0$ in large samples. ($Z^2$ is often reported and follows a $\chi^2_1$ distribution under the null.)

A likelihood ratio test statistic has somewhat better properties for

comparing nested models. We take

$$2\left[\log(L(\text{larger})) - \log(L(\text{smaller}))\right]$$

or

$$(-2\log(L(\text{smaller}))) - (-2\log(L(\text{larger})))$$

and compare the resulting value to a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the two models.

For small samples, these tests should not be used, and exact methods should be used instead. You can learn more about exact methods in BIOS665 next fall.

## Example: French Skiers

Below is the SAS code used to analyze the French skier data.

```
data ski;
input vitc cold count;
cards;
1 1 17
1 0 122
0 1 31
0 0 109
;

proc logistic descending;
freq count;
model cold=vitc;
run;
```

We use the DESCENDING option to request that the response value ordering be reversed. Thus we wish to model $\Pr(\text{cold}) = \Pr(y = 1)$ instead of $\Pr(y = 0)$. We use the FREQ COUNT statement to tell SAS that we have entered the data in a summary form. We would eliminate

this statement if we had entered 17 lines coded "1 1" for vitc and cold, 122 lines coded "1 0" for vitc and cold, 31 lines coded "0 1" for vitc and cold, and 109 lines coded "0 0" for vitc and cold. Since we do not have one line per subject in this data, we need to let SAS know this.

Selected SAS output is provided below.

```
                    The LOGISTIC Procedure

                      Model Information

Data Set                      WORK.SKI
Response Variable             cold
Number of Response Levels     2
Number of Observations        4
Frequency Variable            count
Sum of Frequencies            279
Model                         binary logit
Optimization Technique        Fishers scoring


                     Response Profile

      Ordered                      Total
        Value          cold     Frequency
```

```
                1              1              48
                2              0             231

          Probability modeled is cold=1.

             Model Convergence Status

     Convergence criterion (GCONV=1E-8) satisfied.

               Model Fit Statistics

                                        Intercept
                        Intercept          and
   Criterion              Only        Covariates

   AIC                  258.184         255.312
   SC                   261.815         262.575
   -2 Log L             256.184         251.312
```

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -1.2574 | 0.2035 | 38.1575 | <.0001 |
| vitc | 1 | -0.7134 | 0.3293 | 4.6934 | 0.0303 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| vitc | 0.490 | 0.257 | 0.934 |

From the Wald test and the resulting confidence limits, we see evidence that Vitamin C is protective. Note: The confidence interval for the odds ratio was computed as $(\exp(-0.7134 - 1.96(0.3293)), \exp(-0.7134 + 1.96(0.3293)))$. Other methods for computing this confidence interval may be preferred.

## Example: PIN Study

In the PIN study, we are interested in whether the probability of a preterm birth (before 37 weeks) is affected by a variety of covariates, including

- PTBANY, which equals 1 if the woman has delivered a previous preterm infant and 0 otherwise,
- RACEIND, which equals 1 if the woman is African American and 0 otherwise,
- SMOKEIND, which equals 1 for smokers and 0 otherwise,
- BMIIND, which equals 1 for underweight women and 0 otherwise, and
- EDIND, which equals 1 for women without a high school diploma and 0 otherwise.

We will fit the model

$$\begin{aligned}\text{logit}(\text{Pr}(\text{preterm})) &= \beta_0 + \beta_1 PTBANY + \beta_2 RACEIND \\ &\quad + \beta_3 SMOKEIND + \beta_4 BMIIND \\ &\quad + \beta_5 EDIND.\end{aligned}$$

```
proc logistic descending;
model c_case1=ptbany raceind smokeind bmiind edind;
run;
*****************************************************************
                          The LOGISTIC Procedure

                            Model Information

                Data Set                      WORK.NEW
                Response Variable             c_case1
                Number of Response Levels     2
                Number of Observations        3093
                Model                         binary logit
                Optimization Technique        Fishers scoring


                           Response Profile

                 Ordered                        Total
                   Value       c_case1      Frequency

                       1             1            397
```

```
                    2            0          2696

                Probability modeled is c_case1=1.

NOTE: 633 observations were deleted due to missing values for the response or
      explanatory variables.


                    Model Convergence Status

          Convergence criterion (GCONV=1E-8) satisfied.


                     Model Fit Statistics

                                          Intercept
                           Intercept         and
          Criterion          Only        Covariates

          AIC              1938.403       1888.588
          SC               1944.239       1923.604
          -2 Log L         1936.403       1876.588
```

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 59.8152 | 5 | <.0001 |
| Score | 70.2530 | 5 | <.0001 |
| Wald | 65.3509 | 5 | <.0001 |

The SAS System 2

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.2645 | 0.1073 | 445.0116 | <.0001 |
| ptbany | 1 | 1.0798 | 0.1402 | 59.2982 | <.0001 |

| | | | | | |
|---|---|---|---|---|---|
| raceind | 1 | 0.1944 | 0.1255 | 2.3996 | 0.1214 |
| smokeind | 1 | 0.1928 | 0.3201 | 0.3628 | 0.5469 |
| bmiind | 1 | -0.1170 | 0.1694 | 0.4774 | 0.4896 |
| edind | 1 | 0.1346 | 0.1244 | 1.1706 | 0.2793 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| ptbany | 2.944 | 2.237 | 3.876 |
| raceind | 1.215 | 0.950 | 1.553 |
| smokeind | 1.213 | 0.648 | 2.271 |
| bmiind | 0.890 | 0.638 | 1.240 |
| edind | 1.144 | 0.897 | 1.460 |

1. What is the probability of preterm birth in the sample corresponding to the reference level of all categorical predictors in the model?

2. Are women with prior preterm deliveries at higher or lower risk? Is this risk significantly different from women without a history of preterm delivery?

3. Are African American women at higher or lower risk than women of other ethnicities?

4. What is the relationship between smoking and preterm risk?

5. Is being underweight protective or possibly harmful?

6. What is the effect of education on the outcome?

7. What is the model-predicted probability of preterm birth for an African American woman with a PhD in biostatistics who is a non-smoker, has not had previous children, and has a normal BMI?

Now, suppose we wish to test whether a model with prior preterm delivery as the only predictor is sufficient. We fit this model below.

```
proc logistic descending;
model c_case1=ptbany;
run;
*****************************************************************************

                          The LOGISTIC Procedure

                             Model Information

                 Data Set                     WORK.NEW
                 Response Variable            c_case1
                 Number of Response Levels    2
                 Number of Observations       3093
                 Model                        binary logit
                 Optimization Technique       Fishers scoring


                           Response Profile
```

| Ordered Value | c_case1 | Total Frequency |
|---|---|---|
| 1 | 1 | 397 |
| 2 | 0 | 2696 |

Probability modeled is c_case1=1.

NOTE: 70 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

| Criterion | Intercept Only | Intercept and Covariates |
|---|---|---|

| | | |
|---|---|---|
| AIC | 2372.762 | 2314.908 |
| SC | 2378.799 | 2326.982 |
| -2 Log L | 2370.762 | 2310.908 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 59.8536 | 1 | <.0001 |
| Score | 71.0827 | 1 | <.0001 |
| Wald | 66.5247 | 1 | <.0001 |

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Intercept | 1 | -2.1093 | 0.0623 | 1146.6468 | <.0001 |
| ptbany | 1 | 1.0421 | 0.1278 | 66.5247 | <.0001 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| ptbany | 2.835 | 2.207 | 3.642 |

What do you conclude about the sufficiency of the smaller model?

## Interaction in Logistic Regression

Consider a study on urinary tract infections (UTI) (Koch et al., 1985). Patients were classified as having either complicated (more difficult to cure) or uncomplicated diagnosis of UTI. Because the complicated cases are more difficult to cure, investigators are interested in whether the diagnostic status of the UTI affected the effectiveness of treatment. This hypothesis corresponds to a treatment by diagnosis interaction. In this study, three treatments (A, B, and C) were provided to patients with UTI. The data are given in the table below.

| Diagnosis | Treatment | Cured | Not Cured | % Cured |
|---|---|---|---|---|
| Complicated | A | 78 | 28 | 0.74 |
| Complicated | B | 101 | 11 | 0.90 |
| Complicated | C | 68 | 46 | 0.60 |
| Uncomplicated | A | 40 | 5 | 0.89 |
| Uncomplicated | B | 54 | 5 | 0.92 |
| Uncomplicated | C | 34 | 6 | 0.85 |

We consider the model

$$\begin{aligned}\text{logit}(p) &= \beta_0 + \beta_1 I(\text{TRTA}) + \beta_2 I(\text{TRTB}) \\ &\quad + \beta_3 I(\text{COMP}) + \beta_4 I(\text{TRTA, COMP}) \\ &\quad + \beta_5 I(\text{TRTB, COMP}),\end{aligned}$$

where $I(\cdot)$ represents an indicator variable.

That is,

$$I(E) = \begin{cases} 1 & \text{if E is true} \\ 0 & \text{otherwise} \end{cases}.$$

In this model, the reference group is patients on treatment C with uncomplicated diagnoses.

Thus the essence of this design is given by

$$
\begin{bmatrix}
\text{logit}(Pr(\text{Cure} \mid \text{A, COMP})) \\
\text{logit}(Pr(\text{Cure} \mid \text{B, COMP})) \\
\text{logit}(Pr(\text{Cure} \mid \text{C, COMP})) \\
\text{logit}(Pr(\text{Cure} \mid \text{A, UNCOMP})) \\
\text{logit}(Pr(\text{Cure} \mid \text{B, UNCOMP})) \\
\text{logit}(Pr(\text{Cure} \mid \text{C, UNCOMP}))
\end{bmatrix}
=
\begin{bmatrix}
\beta_0 + \beta_1 + \beta_3 + \beta_4 \\
\beta_0 + \beta_2 + \beta_3 + \beta_5 \\
\beta_0 + \beta_3 \\
\beta_0 + \beta_1 \\
\beta_0 + \beta_2 \\
\beta_0
\end{bmatrix}
$$

$$
= \begin{bmatrix}
1 & 1 & 0 & 1 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5
\end{bmatrix}.
$$

Below is the SAS code used to fit the interaction model as well as a reduced model with no diagnosis by treatment interaction.

```
data uti;
input diagnosis $ trt $ cure $ count;
cards;
complicated A cured 78
complicated A not 28
complicated B cured 101
complicated B not 11
complicated C cured 68
complicated C not 46
uncomplicated A cured 40
uncomplicated A not 5
uncomplicated B cured 54
uncomplicated B not 5
uncomplicated C cured 34
uncomplicated C not 6
;

proc logistic;
freq count;
```

```
class diagnosis trt/param=ref;
model cure=diagnosis|trt;
run;

proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt;
run;
```

The SAS output is provided below.

```
                    The SAS System                          1

                 The LOGISTIC Procedure

                   Model Information

Data Set                      WORK.UTI
Response Variable             cure
Number of Response Levels     2
Number of Observations        12
Frequency Variable            count
Sum of Frequencies            476
Model                         binary logit
Optimization Technique        Fishers scoring

                   Response Profile

      Ordered                       Total
        Value     cure          Frequency
```

```
        1     cured                375
        2     not                  101

Probability modeled is cure=cured.

     Class Level Information

                                 Design
                                Variables

Class          Value            1      2

diagnosis      complica         1
               uncompli         0

trt            A                1      0
               B                0      1
               C                0      0


      Model Convergence Status
```

```
Convergence criterion (GCONV=1E-8) satisfied.


                    Model Fit Statistics

                                          Intercept
                          Intercept          and
      Criterion             Only        Covariates

      AIC                  494.029        459.556
      SC                   498.194        484.549
      -2 Log L             492.029        447.556

                    The SAS System                                  2

                    The LOGISTIC Procedure

      Testing Global Null Hypothesis: BETA=0

Test                      Chi-Square       DF     Pr > ChiSq
```

| | | | |
|---|---|---|---|
| Likelihood Ratio | 44.4726 | 5 | <.0001 |
| Score | 44.7864 | 5 | <.0001 |
| Wald | 39.9312 | 5 | <.0001 |

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| diagnosis | 1 | 7.7653 | 0.0053 |
| trt | 2 | 1.0069 | 0.6045 |
| diagnosis*trt | 2 | 2.6384 | 0.2674 |

Analysis of Maximum Likelihood Estimates

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| Intercept | | 1 | 1.7346 | 0.4428 | 15.3451 | <.0001 |
| diagnosis | complica | 1 | -1.3437 | 0.4822 | 7.7653 | 0.0053 |

| | | | | | | |
|---|---|---|---|---|---|---|
| trt | A | 1 | 0.3448 | 0.6489 | 0.2824 | 0.5952 |
| trt | B | 1 | 0.6445 | 0.6438 | 1.0020 | 0.3168 |
| diagnosis*trt | complica A | 1 | 0.2888 | 0.7114 | 0.1649 | 0.6847 |
| diagnosis*trt | complica B | 1 | 1.1818 | 0.7428 | 2.5311 | 0.1116 |

The SAS System 3

The LOGISTIC Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.UTI |
| Response Variable | cure |
| Number of Response Levels | 2 |
| Number of Observations | 12 |
| Frequency Variable | count |
| Sum of Frequencies | 476 |
| Model | binary logit |
| Optimization Technique | Fishers scoring |

Response Profile

```
 Ordered                          Total
   Value     cure             Frequency

       1     cured                  375
       2     not                    101

Probability modeled is cure=cured.

      Class Level Information

                                 Design
                               Variables

Class         Value             1      2

diagnosis     complica          1
              uncompli          0

trt           A                 1      0
              B                 0      1
              C                 0      0
```

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

| Criterion | Intercept Only | Intercept and Covariates |
|---|---|---|
| AIC | 494.029 | 458.071 |
| SC | 498.194 | 474.733 |
| -2 Log L | 492.029 | 450.071 |

The SAS System 4

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 41.9579 | 3 | <.0001 |
| Score | 38.8456 | 3 | <.0001 |
| Wald | 34.9484 | 3 | <.0001 |

Type III Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| diagnosis | 1 | 10.2885 | 0.0013 |
| trt | 2 | 24.6219 | <.0001 |

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|

| | | | | | | |
|---|---|---|---|---|---|---|
| Intercept | | 1 | 1.4184 | 0.2987 | 22.5505 | <.0001 |
| diagnosis | complica | 1 | -0.9616 | 0.2998 | 10.2885 | 0.0013 |
| trt | A | 1 | 0.5847 | 0.2641 | 4.9020 | 0.0268 |
| trt | B | 1 | 1.5608 | 0.3160 | 24.4010 | <.0001 |

Odds Ratio Estimates

| Effect | | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|---|
| diagnosis | complica vs uncompli | 0.382 | 0.212 | 0.688 |
| trt | A vs C | 1.795 | 1.069 | 3.011 |
| trt | B vs C | 4.762 | 2.564 | 8.847 |

To test the significance of the interaction effect, we conduct a likelihood ratio test of the full versus reduced models. The null hypothesis is given by $H_0 : \beta_4 = \beta_5 = 0$. The difference between 447.556 (interaction) and 450.071 (no interaction) is 2.515, which we compare to the $\chi^2_2$ distribution because we are testing two interaction terms. The critical value for this distribution is 5.99, so we cannot reject the null hypothesis, and we conclude that the interaction between treatment and diagnosis is not significant.

We see that cure is much less likely for patients with a complicated diagnosis. In addition, both treatments A and B are superior to treatment C.

We cannot tell from the output provided whether treatment B is significantly better than treatment A. In order to test this hypothesis, we need to construct a contrast. The hypothesis of interest is $H_0 : \beta_1 = \beta_2$ and is tested using the following SAS code.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt;
contrast B vs. A trt -1 1/estimate=exp;
/* ESTIMATE=EXP option requests that the OR is printed */
run;
```

The additional SAS output is provided below.

Contrast Test Results

| Contrast | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| B vs. A | 1 | 8.6919 | 0.0032 |

The SAS System

The LOGISTIC Procedure

Contrast Rows Estimation and Testing Results

| Contrast | Type | Row | Estimate | Standard Error | Alpha | Lower Limit | Upper Limit |
|---|---|---|---|---|---|---|---|
| B vs. A | EXP | 1 | 2.6539 | 0.8786 | 0.05 | 1.3870 | 5.0778 |

```
Contrast Rows Estimation and Testing Results

                                  Wald
Contrast Type      Row Chi-Square    Pr > ChiSq

B vs. A  EXP         1     8.6919        0.0032
```

We see that treatment B is indeed superior to treatment A. Patients on treatment B have 2.65 times higher odds of cure than those on treatment A.

We may also calculate predicted probabilities and odds from the main effects model.

| Diagnosis | Trt | $Pr(\text{Cure}) = p$ | Odds of Cure |
|---|---|---|---|
| Comp | A | $\frac{\exp(\beta_0+\beta_1+\beta_3)}{1+\exp(\beta_0+\beta_1+\beta_3)}$ | $\exp(\beta_0 + \beta_1 + \beta_3)$ |
| Comp | B | $\frac{\exp(\beta_0+\beta_2+\beta_3)}{1+\exp(\beta_0+\beta_2+\beta_3)}$ | $\exp(\beta_0 + \beta_2 + \beta_3)$ |
| Comp | C | $\frac{\exp(\beta_0+\beta_3)}{1+\exp(\beta_0+\beta_3)}$ | $\exp(\beta_0 + \beta_3)$ |
| Uncomp | A | $\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$ | $\exp(\beta_0 + \beta_1)$ |
| Uncomp | B | $\frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$ | $\exp(\beta_0 + \beta_2)$ |
| Uncomp | C | $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ | $\exp(\beta_0)$ |

## Goodness of Fit for Logistic Regression

After fitting a model, we need to know just how well that model fits the data. We measure this by the closeness of the model-predicted values to the corresponding observed values. (Is the predicted probability of cure, $\widehat{p}$, close to 0 for those not cured and close to 1 for those cured?)

*Goodness-of-fit statistics* are test statistics used to assess differences between observed and predicted values, which we expect to be random. If the cell counts are sufficiently large, then these test statistics approximately follow chi-squared distributions.

Consider the following contingency table.

| Diagnosis | Treatment | Cured $(j=1)$ | Not Cured $(j=2)$ |
|---|---|---|---|
| Comp $(h=1)$ | A $(i=1)$ | $n_{111}$ | $n_{112}$ |
| Comp $(h=1)$ | B $(i=2)$ | $n_{121}$ | $n_{122}$ |
| Comp $(h=1)$ | C $(i=3)$ | $n_{131}$ | $n_{132}$ |
| Unoomp $(h=2)$ | A $(i=1)$ | $n_{211}$ | $n_{212}$ |
| Unoomp $(h=2)$ | B $(i=2)$ | $n_{221}$ | $n_{222}$ |
| Unoomp $(h=2)$ | C $(i=3)$ | $n_{231}$ | $n_{232}$ |

The *Pearson chi-square goodness-of-fit test* compares observed and model-predicted cell counts by computing

$$Q_p = \sum_{h=1}^{2} \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(n_{hij} - m_{hij})^2}{m_{hij}},$$

where $n_{hij}$ are the observed cell counts, and $m_{hij}$ are the model-predicted counts.

We obtain the model-predicted counts by taking

$$m_{hij} = \begin{cases} n_{hi\cdot}\widehat{p}_{hi} & \text{for } j = 1 \\ n_{hi\cdot}(1 - \widehat{p}_{hi}) & \text{for } j = 2 \end{cases}.$$

So to obtain the predicted number of patients cured for level $h$ of diagnosis and level $i$ of treatment, we multiply the total number of subjects with diagnosis $h$ and treatment $i$, which is $n_{hi\cdot}$, by the model-predicted probability of cure for such subjects, given by $\widehat{p}_{hi}$.

These counts may also be obtained in PROC LOGISTIC with the inclusion of the statement OUTPUT OUT=PREDICT PRED=PROB (and then printing the data in PREDICT).

For example, for subjects with a complicated diagnosis on treatment A, we observed 78 cures and 28 failures. Using the model

$$\begin{aligned}\text{logit}(p) &= \beta_0 + \beta_1 I(\text{TRTA}) + \beta_2 I(\text{TRTB}) \\ &\quad + \beta_3 I(\text{COMP}),\end{aligned}$$

we obtained $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1.4184 & 0.5847 & 1.5608 & -0.9616 \end{pmatrix}'$.

Thus we compute the expected number

$$\begin{aligned} m_{111} &= \begin{cases} 106 \ \frac{\exp(1.4184+0.5847-0.9616)}{1+\exp(1.4184+0.5847-0.9616)} & \text{cured} \\ 106 \ \left(1 - \frac{\exp(1.4184+0.5847-0.9616)}{1+\exp(1.4184+0.5847-0.9616)}\right) & \text{not cured} \end{cases} \\ &= \begin{cases} 106(0.74) = 78.35 & \text{expected patients cured} \\ 106(0.26) = 27.65 & \text{expected patients not cured} \end{cases}. \end{aligned}$$

The Pearson chi-squared statistic can tell us whether the fit is sufficient.

We use the following SAS code to compute Pearson's chi-squared statistic for the UTI data.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt/scale=none aggregate;
/* SCALE produces goodness-of-fit statistics */
/* AGGREGATE tells LOGISTIC to treat each unique combination of */
/* explanatory variables as a distinct group in computing the GOF stats */
run;
*****************************************************************
          Deviance and Pearson Goodness-of-Fit Statistics

    Criterion        DF          Value     Value/DF     Pr > ChiSq
    Deviance          2         2.5147       1.2573         0.2844
    Pearson           2         2.7574       1.3787         0.2519
                  Number of unique profiles: 6
```

The non-significance of this chi-squared test means that the model fits adequately.

The *deviance* or likelihood ratio chi-square, $Q_L$, is another traditional measure of goodness-of-fit, which is given by

$$Q_L = \sum_{h=1}^{2} \sum_{i=1}^{3} \sum_{j=1}^{2} 2n_{hij} \log \left( \frac{n_{hij}}{m_{hij}} \right).$$

We interpret this test in the same way as the Pearson chi-squared test. (Note that the ratios will be close to one, and thus their logarithms close to zero, for a model that fits well.)

These two goodness of fit tests are valid only for large enough samples. We need for each group $n_{hi\cdot}$ to have at least 10 subjects, we need 80% of the predicted counts $m_{hij}$ to be at least 5, and all other expected counts to be greater than 2 in order for these tests to be valid. If these conditions do not hold, exact methods for logistic regression are available.

These goodness-of-fit statistics are calculated assuming all predictors are categorical. However, continuous predictors are commonly used in logistic regression models. In such a case, sample size requirements for the validity of the above goodness-of-fit tests will rarely be met! Alternatively, fit can be assessed by fitting an appropriate expanded model with additional explanatory variables (including interactions if desired) and examining the difference in the log-likelihoods using a likelihood ratio test. Another strategy is to consider the Hosmer and Lemeshow goodness-of-fit statistic. This test places subjects into deciles based on model-predicted probabilities and then computes a Pearson chi-squared test based on the observed and expected number of subjects in the deciles. The statistic is then compared to a chi-squared distribution and is available by specifying the LACKFIT option in the MODEL statement.

## Diagnostics for Logistic Regression

Although goodness-of-fit statistics tell you how well the model fits the data, they do not tell you much about where a particular model fails to fit the data, or *lack of fit*.

Suppose that you have $s$ groups, $i = 1, \ldots, s$, with $n_i$ subjects in group $i$ and $y_i$ events in group $i$. *Pearson residuals* (the sum of their squares is $Q_P$, the Pearson chi-square goodness-of-fit statistic) are given by

$$e_i = \frac{y_i - n_i \widehat{p}_i}{\sqrt{n_i \widehat{p}_i (1 - \widehat{p}_i)}}.$$

These residuals compare the differences between observed counts and their predicted values, scaled by the observed count's standard deviation. Generally, residuals are considered indicative of lack of fit if they exceed 2 in absolute value. (Note that the $e_i$ are formed by subtracting the mean and standardizing by the square root of the variance for binomial data.)

*Deviance residuals* (the sum of their squares yields the deviance statistic) are given by

$$d_i = \mathsf{sgn}(y_i - \widehat{y}_i)\left[2y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \widehat{y}_i}\right)\right]^{\frac{1}{2}},$$

where $\widehat{y}_i = n\widehat{p}_i$.

The SAS code and output for calculating these residuals is provided below.

```
proc logistic;
freq count;
class diagnosis trt/param=ref;
model cure=diagnosis trt/influence;
run;
*****************************************************************
                        Regression Diagnostics

                                Covariates
            Case
          Number  diagnosiscomplica         trtA         trtB

               1             1.0000       1.0000            0
               2             1.0000       1.0000            0
               3             1.0000            0       1.0000
               4             1.0000            0       1.0000
               5             1.0000            0            0
               6             1.0000            0            0
               7                  0       1.0000            0
```

| | | | |
|---|---|---|---|
| 8 | 0 | 1.0000 | 0 |
| 9 | 0 | 0 | 1.0000 |
| 10 | 0 | 0 | 1.0000 |
| 11 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 |

Regression Diagnostics

```
                 Pearson Residual                 Deviance Residual

  Case                (1 unit = 0.55)                  (1 unit = 0.31)
 Number     Value   -8  -4  0 2 4 6 8      Value   -8  -4  0 2 4 6 8

      1    0.5941  |        |*       |    0.7775  |        |  *     |
      2   -1.6833  |     *  |        |   -1.6394  |   *    |        |
      3    0.3647  |        |*       |    0.4997  |        | *      |
      4   -2.7422  |   *    |        |   -2.0700  | *      |        |
      5    0.7958  |        |*       |    0.9906  |        |  *     |
      6   -1.2566  |      * |        |   -1.3765  |    *   |        |
      7    0.3673  |        |*       |    0.5031  |        | *      |
      8   -2.7226  |   *    |        |   -2.0638  | *      |        |
      9    0.2255  |        *        |    0.3149  |        |*       |
```

```
10   -4.4352  |*       |        |   -2.4612  |*       |        |
11    0.4920  |        |*       |    0.6585  |        | *      |
12   -2.0324  |    *   |        |   -1.8084  |  *     |        |
```

Based on the Pearson and deviance residuals, we see that the model fits poorly for quite a few groups, with the worst fit for uncured uncomplicated diagnoses with treatment B. Perhaps an interaction model (or the addition of other covariates) would improve the model fit. Although it seems that we may have conflicting results from examination of the Pearson residuals and the Pearson chi-squared statistic, we see that while the model does fit the data, there is some suggestion that we can still do better!

## Example: Cirrhosis Data

The Mayo Clinic conducted a double-blinded randomized clinical trial in patients with primary biliary cirrhosis (PBC) of the liver to compare the drug D-penicillamine (DPCA) with placebo. PBC is a rare but fatal chronic liver disease (in the years since the trial was completed, the disease has become more treatable due to advances in liver transplantation).

We consider status (1=death, 0=otherwise) as the outcome of interest in comparing the two treatments. In addition, we have information about a variety of covariates, including the following.

- drug (1=DPCA, 0=placebo)
- age in days
- sex (0=male, 1=female)
- presence of ascites (0=no 1=yes)

- presence of hepatomegaly (0=no 1=yes)
- presence of spiders (0=no 1=yes)
- presence of edema (0=no edema and no diuretic therapy for edema; .5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy)
- serum bilirubin in mg/dl
- serum cholesterol in mg/dl
- albumin in gm/dl
- urine copper in ug/day
- alkaline phosphatase in U/liter
- SGOT in U/ml
- triglicerides in mg/dl
- platelets per cubic ml / 1000

- prothrombin time in seconds
- histologic stage of disease

First, we do some data checking.

```
proc means;
var age bili chol albumin copper alk_phos sgot trig platelet protime;
run;

proc freq;
tables status drug sex ascites hepatom spiders edema stage;
run;
```

```
*************************************************************
                          The SAS System                    1

                       The MEANS Procedure

Variable    N          Mean       Std Dev       Minimum       Maximum
-------------------------------------------------------------------
age       312      18269.44       3864.81       9598.00      28650.00
bili      312     3.2560897     4.5303153     0.3000000    28.0000000
chol      284   369.5105634   231.9445450   120.0000000       1775.00
albumin   312     3.5200000     0.4198920     1.9600000     4.6400000
copper    310    97.6483871    85.6139199     4.0000000   588.0000000
alk_phos  312       1982.66       2140.39   289.0000000      13862.40
```

```
sgot      312   122.5563462    56.6995249    26.3500000   457.2500000
trig      282   124.7021277    65.1486387    33.0000000   598.0000000
platelet  308   261.9350649    95.6087423    62.0000000   563.0000000
protime   312    10.7256410     1.0043232     9.0000000    17.1000000
-----------------------------------------------------------------------

                          The SAS System                          2

                        The FREQ Procedure

                                          Cumulative    Cumulative
      status    Frequency     Percent     Frequency      Percent
      -----------------------------------------------------------
         0         187       59.94           187        59.94
         1         125       40.06           312       100.00


                                          Cumulative    Cumulative
      drug    Frequency     Percent     Frequency      Percent
      -----------------------------------------------------------
         0         154       49.36           154        49.36
         1         158       50.64           312       100.00
```

| sex | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 36 | 11.54 | 36 | 11.54 |
| 1 | 276 | 88.46 | 312 | 100.00 |

| ascites | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 288 | 92.31 | 288 | 92.31 |
| 1 | 24 | 7.69 | 312 | 100.00 |

| hepatom | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 152 | 48.72 | 152 | 48.72 |
| 1 | 160 | 51.28 | 312 | 100.00 |

| spiders | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 222 | 71.15 | 222 | 71.15 |
| 1 | 90 | 28.85 | 312 | 100.00 |

| edema | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 263 | 84.29 | 263 | 84.29 |
| 0.5 | 29 | 9.29 | 292 | 93.59 |
| 1 | 20 | 6.41 | 312 | 100.00 |

The SAS System 3

The FREQ Procedure

| stage | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|

```
-------------------------------------------------------
    1          16        5.13            16         5.13
    2          67       21.47            83        26.60
    3         120       38.46           203        65.06
    4         109       34.94           312       100.00
```

Based on this checking, we will convert age into years and pseudo-center at age 50. In addition, we will convert alkaline phosphate into U/kL (divide by 1000). We will also create two edema variables: EDEMAD for those without relief from diuretics, and EDEMAND for the patients with milder cases. We will also create four variables for disease stage: STAGE1, STAGE2, STAGE3, and STAGE4. The SAS code for making these changes is presented below.

```
data pbc;
set pbc;
age=age/365.25-50;
alk_phos=alk_phos/1000;
edema_d=0;
```

```
edema_nd=0;
if edema=1 then edema_d=1;
if edema=.5 then edema_nd=1;
stage1=0; stage2=0; stage3=0; stage4=0;
if stage=1 then stage1=1;
if stage=2 then stage2=1;
if stage=3 then stage3=1;
if stage=4 then stage4=1;
run;
```

First, we fit a large model including all the predictors.

```
proc logistic descending;
model status=drug sex ascites hepatom spiders edema_d edema_nd stage2
        stage3 stage4 age bili chol albumin copper alk_phos sgot trig
        platelet protime;
run;
*************************************************************
                          The SAS System                    1

                      The LOGISTIC Procedure

                        Model Information

            Data Set                      WORK.PBC
            Response Variable             status
            Number of Response Levels     2
            Number of Observations        276
            Model                         binary logit
            Optimization Technique        Fishers scoring
```

```
                    Response Profile

          Ordered                        Total
            Value          status     Frequency

                1               1           111
                2               0           165

          Probability modeled is status=1.


             Model Convergence Status

     Convergence criterion (GCONV=1E-8) satisfied.


              Model Fit Statistics

                                       Intercept
                        Intercept         and
        Criterion         Only        Covariates
```

| | | |
|---|---|---|
| AIC | 373.984 | 277.117 |
| SC | 377.604 | 353.145 |
| -2 Log L | 371.984 | 235.117 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 136.8671 | 20 | <.0001 |
| Score | 108.2980 | 20 | <.0001 |
| Wald | 64.5753 | 20 | <.0001 |

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -12.7956 | 3.8279 | 11.1739 | 0.0008 |

| | | | | | |
|---|---|---|---|---|---|
| drug | 1 | 0.3091 | 0.3413 | 0.8201 | 0.3652 |
| sex | 1 | -0.6448 | 0.5355 | 1.4500 | 0.2285 |
| ascites | 1 | 1.0535 | 1.3898 | 0.5746 | 0.4484 |
| hepatom | 1 | 0.1144 | 0.4000 | 0.0818 | 0.7749 |
| spiders | 1 | 0.3533 | 0.4001 | 0.7797 | 0.3772 |
| edema_d | 1 | 0.7758 | 1.4717 | 0.2779 | 0.5981 |
| edema_nd | 1 | -0.0705 | 0.6022 | 0.0137 | 0.9068 |
| stage2 | 1 | 2.5956 | 1.4906 | 3.0323 | 0.0816 |
| stage3 | 1 | 2.8408 | 1.4747 | 3.7106 | 0.0541 |
| stage4 | 1 | 2.8148 | 1.4855 | 3.5905 | 0.0581 |
| age | 1 | 0.0519 | 0.0183 | 8.0751 | 0.0045 |
| bili | 1 | 0.1532 | 0.0835 | 3.3689 | 0.0664 |
| chol | 1 | 0.000322 | 0.000857 | 0.1407 | 0.7076 |
| albumin | 1 | -0.1408 | 0.5006 | 0.0791 | 0.7786 |
| copper | 1 | 0.00285 | 0.00250 | 1.3027 | 0.2537 |
| alk_phos | 1 | 0.2708 | 0.0898 | 9.0973 | 0.0026 |
| sgot | 1 | 0.00584 | 0.00322 | 3.2853 | 0.0699 |
| trig | 1 | 0.00248 | 0.00330 | 0.5644 | 0.4525 |
| platelet | 1 | -0.00003 | 0.00199 | 0.0002 | 0.9880 |
| protime | 1 | 0.7345 | 0.2143 | 11.7498 | 0.0006 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| drug | 1.362 | 0.698 | 2.659 |
| sex | 0.525 | 0.184 | 1.499 |
| ascites | 2.868 | 0.188 | 43.704 |
| hepatom | 1.121 | 0.512 | 2.455 |
| spiders | 1.424 | 0.650 | 3.119 |
| edema_d | 2.172 | 0.121 | 38.873 |
| edema_nd | 0.932 | 0.286 | 3.034 |
| stage2 | 13.405 | 0.722 | 248.900 |
| stage3 | 17.129 | 0.952 | 308.348 |
| stage4 | 16.690 | 0.908 | 306.819 |
| age | 1.053 | 1.016 | 1.092 |
| bili | 1.166 | 0.990 | 1.373 |
| chol | 1.000 | 0.999 | 1.002 |
| albumin | 0.869 | 0.326 | 2.317 |
| copper | 1.003 | 0.998 | 1.008 |
| alk_phos | 1.311 | 1.099 | 1.563 |
| sgot | 1.006 | 1.000 | 1.012 |

```
trig          1.002      0.996      1.009
platelet      1.000      0.996      1.004
protime       2.085      1.370      3.173
```

Suppose that based on scientific reasoning supplied in advance by the investigator, a reduced model was fit to the data. This model did not contain the variables hepatom, edema, chol, albumin, sgot, trig, or platelet.

```
proc logistic descending;
model status=drug sex ascites spiders stage2 stage3 stage4 age bili copper alk_phos
run;
*****************************************************************
                        The LOGISTIC Procedure

                          Model Information

            Data Set                      WORK.PBC
            Response Variable             status
            Number of Response Levels     2
            Number of Observations        276
            Model                         binary logit
```

```
Optimization Technique        Fishers scoring


                Response Profile

      Ordered                        Total
        Value          status     Frequency

            1               1           111
            2               0           165

        Probability modeled is status=1.


           Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.


            Model Fit Statistics

                                   Intercept
```

| Criterion | Intercept Only | and Covariates |
|---|---|---|
| AIC | 373.984 | 265.949 |
| SC | 377.604 | 313.014 |
| -2 Log L | 371.984 | 239.949 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 132.0346 | 12 | <.0001 |
| Score | 104.5750 | 12 | <.0001 |
| Wald | 62.6444 | 12 | <.0001 |

The SAS System 5

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -11.5823 | 3.0107 | 14.8000 | 0.0001 |
| drug | 1 | 0.2729 | 0.3327 | 0.6732 | 0.4119 |
| sex | 1 | -0.6684 | 0.5141 | 1.6903 | 0.1936 |
| ascites | 1 | 1.3471 | 1.1475 | 1.3782 | 0.2404 |
| spiders | 1 | 0.3759 | 0.3805 | 0.9757 | 0.3233 |
| stage2 | 1 | 2.6314 | 1.4146 | 3.4605 | 0.0629 |
| stage3 | 1 | 2.9260 | 1.3958 | 4.3941 | 0.0361 |
| stage4 | 1 | 2.8652 | 1.3750 | 4.3424 | 0.0372 |
| age | 1 | 0.0461 | 0.0170 | 7.3783 | 0.0066 |
| bili | 1 | 0.2330 | 0.0727 | 10.2842 | 0.0013 |
| copper | 1 | 0.00301 | 0.00247 | 1.4928 | 0.2218 |
| alk_phos | 1 | 0.3000 | 0.0874 | 11.7794 | 0.0006 |
| protime | 1 | 0.6613 | 0.1986 | 11.0904 | 0.0009 |

Odds Ratio Estimates

| | Point | 95% Wald |
|---|---|---|

| Effect | Estimate | Confidence Limits | |
|---|---|---|---|
| drug | 1.314 | 0.685 | 2.522 |
| sex | 0.513 | 0.187 | 1.404 |
| ascites | 3.846 | 0.406 | 36.461 |
| spiders | 1.456 | 0.691 | 3.070 |
| stage2 | 13.894 | 0.868 | 222.276 |
| stage3 | 18.652 | 1.209 | 287.655 |
| stage4 | 17.553 | 1.186 | 259.844 |
| age | 1.047 | 1.013 | 1.083 |
| bili | 1.262 | 1.095 | 1.456 |
| copper | 1.003 | 0.998 | 1.008 |
| alk_phos | 1.350 | 1.137 | 1.602 |
| protime | 1.937 | 1.313 | 2.859 |

To test whether the reduced model is sufficient, we may compare log-likelihoods:
239.949 (smaller) - 235.117 (larger)=4.832. Comparing this to a $\chi^2_8$ random variable (with critical value 15.51), we conclude that the reduced model is sufficient.

Although we could reduce the model further, the investigator wanted to leave the remaining terms in the model for scientific reasons (to show he/she adjusted for the other factors).

1. What effect does treatment have on survival?

2. What factors work to lengthen survival? How do you interpret the benefit of these factors?

3. What factors are associated with diminished survival? How do you interpret their effects?

# Lecture 19: Poisson Regression

## Poisson Regression

Recall that a random variable y has Poisson distribution if

$$Pr(y = k) = \frac{\mu^k}{k!} e^{-\mu}, \mu > 0, k = 1, 2, \cdot.s$$

We have $E(y) = var(y) = \mu > 0$. This distribution is used for counts of events occur randomly over time or space, when outcomes in disjoint periods or regions are independent. Examples are plentiful such as traffic accidents, the incidence of rare events or diseases, etc.

**African Elephants Data** The first of our examples deals with the number of successful matings of 41 male African elephants over a period of eight years, and to examine to what extent these numbers depend on their ages at the onset of the study.

| Age | Matings | Age | Matings | Age | Matings |
|---|---|---|---|---|---|
| 27 | 0 | 33 | 3 | 39 | 1 |
| 28 | 1 | 33 | 3 | 41 | 3 |
| 28 | 1 | 33 | 3 | 42 | 4 |
| 28 | 1 | 33 | 2 | 43 | 0 |
| 28 | 3 | 34 | 1 | 43 | 2 |
| 29 | 0 | 34 | 1 | 43 | 3 |
| 29 | 0 | 34 | 2 | 43 | 4 |
| 29 | 0 | 34 | 3 | 43 | 9 |
| 29 | 2 | 36 | 5 | 44 | 3 |
| 29 | 2 | 36 | 6 | 45 | 5 |
| 29 | 2 | 37 | 1 | 47 | 7 |
| 30 | 1 | 37 | 1 | 48 | 2 |

```
32 2 37 6 52 9
33 4 38 2
```

**Poisson Regression** In view of these count data, we will model the conditional distribution of the response Y given the explanatory variables X using the Poisson distribution whose mean is given by $\mu = \mu(\mathbf{x}) = E(y \mid \mathbf{x}) \geq 0$. To ensure this constraint, we assume

$$\log \mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{\beta}^T \mathbf{x}$$

This is called a **Poisson loglinear model**. The coefficient, say $\beta_1$ of $x_1 =$ age, can be interpreted as follows: if we increase age by one unit while keeping the remaining variables fixed. This affects the mean $\mu$ of the Poisson response by $100(e^{\beta_1} - 1)\%$.

## Estimation

- We obtain the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ by maximizing the log-likelihood:

$$l(\boldsymbol{\beta}) = \sum_i^n y_i \log \mu_i - \sum_i^n \mu_i$$

  Also, the MLE is asymptotically normal:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \hat{\mathbf{V}})$$

  .

- The MLE of the mean is $\mu(\mathbf{x}) = \exp(\hat{\boldsymbol{\beta}}\mathbf{x})$. In particular, $\hat{\mu}_i = \hat{\mu}(\mathbf{x}_i) = \exp(\hat{\boldsymbol{\beta}}\mathbf{x}_i)$, $i = 1, 2, \cdots, n$.
- The covariance of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{V}}$ is estimated by

$$\widehat{COV}(\hat{\boldsymbol{\beta}}) = (\sum_i \hat{\mu}_i \mathbf{x}_i \mathbf{x}_i^T)^{-1}$$

  . The $se(\hat{\beta}_j)$ is the square root of the diagonal elements of the

above matrix.

- Approximate 95% CI for the mean are

$$\exp\left(\hat{\boldsymbol{\beta}}\mathbf{x}_i \pm 1.96\sqrt{\mathbf{x}_i^T(\sum_i \hat{\mu}_i\mathbf{x}_i\mathbf{x}_i^T)^{-1}\mathbf{x}_i}\right),$$

  some other ways to calculate the CI might be better.

- The deviance is

$$D = 2\sum_i y_i \log(y_i/\hat{\mu}_i) \approx \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2$$

  , which is the Pearson's chi-square statistic with $n - p - 1$ df. The model for the mean is doubtful when the chi-square statistic is much greater than $n - p - 1$.

- The deviance residuals are defined by

$$d_i = sign(y_i - \hat{\mu}_i)\sqrt{2(y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i))}.$$

Pearson residuals are given by

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}.$$

Elephant Data

```
data a;
infile elephants.txt firstobs=2;
input age matings;

proc genmod;
 model matings = age / dist  = poisson
                       lrci covb;

proc genmod;
 model matings = age age*age / dist  = poisson
                               lrci covb;
```

```
run;
```

**Ceriodaphnia Organisms Data** This data set contains the number of Ceriodaphnia organisms that are counted in a controlled environment in which reproduction occurs among the organisms, and there are two strains involved. The data is located at ceriodaphnia.txt. Ceriodaphnia represents the number of reproduction occurred, Concentration is the chemical component in the environment that impairs the reproduction, and Strain is the strain type.

```
Cerio Conc Strain;
datalines;
82  0.00    1
106 0.00    1
63  0.00    1
45  0.50    1
34  0.50    1
26  0.50    1
11  1.50    1
```

```
10  1.50    1
10  1.50    1
14  1.00    2
14  1.00    2
9   1.25    2
12  1.25    2
16  1.25    2
7   1.50    2
...
```

## Ceriodaphnia Organisms Data

```
proc import datafile = ceriodaphnia.csv
 out = Ceriodaphnia
 dbms = CSV
 REPLACE
 ;
```

```
proc genmod;
model Cerio = Conc Strain / dist  = poisson
                            lrci covb;
run; quit;
```

**Skin Cancer Example**:

- Skin cancer for women in Dallas and Minn
- 8 age group and two cities (0 = Minn, 1 = Dallas)
- $y_{ij}$ = count (number of cases) in age $i$ and city $j$
- $n_{ij}$ = population size in age $i$ and city $j$

**Objective**: Determine whether the risk for skin cancer adjusted for age is higher in one area than in the other.

$risk$ = probability of developing skin cancer
$\lambda_{ij}$ = probability of developing skin cancer in $(i, j)$th group.

| y | city | age | pop |
|---|---|---|---|
| 1 | 0 | 1 | 172675 |
| 16 | 0 | 2 | 123065 |
| 30 | 0 | 3 | 96216 |
| 71 | 0 | 4 | 92051 |
| 102 | 0 | 5 | 72159 |
| 130 | 0 | 6 | 54722 |
| 133 | 0 | 7 | 32185 |
| 40 | 0 | 8 | 8328 |
| 4 | 1 | 1 | 181343 |
| 38 | 1 | 2 | 146207 |
| 119 | 1 | 3 | 121374 |
| 221 | 1 | 4 | 111353 |
| 259 | 1 | 5 | 83004 |
| 310 | 1 | 6 | 55932 |
| 226 | 1 | 7 | 29007 |
| 65 | 1 | 8 | 7538 |

**Poisson Regression** $y_{ij}$ has a poisson distribution with mean

$$E(y_{ij}) = n_{ij}\lambda_{ij}.$$

Poisson regression concerns the model

$$\log(y_{ij}) = \log(n_{ij}) + \beta_0 + \sum_{i=1}^{7} \beta_i age_i + \beta_8 city,$$

which is equivalent to

$$\log(\lambda_{ij}) = \beta_0 + \sum_{i=1}^{7} \beta_i age_i + \beta_8 city.$$

Interpretations:

- $\beta_0 = \log(\lambda_{81}) = \log$ of the risk in age $8$ group in Dallas.
- $\beta_i = \log(\lambda_{i1}) = \log$ of the risk in age $i$ group in Dallas (for $i = 1, 2, ..., 7$).
- $\beta_0 + \beta_8 = \log(\lambda_{80}) = \log$ of the risk in age 8 group in Minn.

- $\beta_i + \beta_8 = \log(\lambda_{i0}) = \log$ of the risk in age $i$ in Minn (for $i = 1, 2, ..., 7$).
- $\beta_8 = \log(\lambda_{i1}) - \log(\lambda_{i0}) = \log(\frac{\lambda_{i1}}{\lambda_{i0}})$
- $RR_i = \frac{\lambda_{i0}}{\lambda_{i1}} = \exp(\beta_8), i = 1, ..., 8$ where $RR_i$ is the relative risk in age $i$ between two cities.

SAS Code

```
*-------------------------------------------------------
---*
| Comparison of incidence of nonmelanoma skin cancer among  |
| women in Minneapolis St. Paul and Dallas Ft Worth.        |
| KKMN, 3e Table 24-1, p688                                 |
*-------------------------------------------------------
---*;
```

```
title LOGISTIC AND POISSON REGRESSION;

proc import datafile = skincancer.csv
 out = skin
 dbms = CSV
 REPLACE
 ;



data skin;
  set skin;
   lpop = log(pop);
run;

proc genmod;
 class age city;
```

```
 model y = age city/ dist  = poisson
 offset= lpop lrci;
 run;


proc genmod;
 class age city;
 model y/pop = age city/ dist  = poisson
                         lrci;
 run;


proc genmod;
 class age city;
 model y/pop = age city/ dist  = bin
                         lrci;
run;
```

Note: For those situations in which $n_{ij}$ is large and $\lambda_{ij}$ is very small, the Poisson distribution can be used to approximate the binomial distribution. The larger the $n_{ij}$ and the smaller the $\lambda_{ij}$, the better is the approximation. For the skin cancer data, we expect $\lambda_{ij}$ to be small, therefore, the two models (Binomial and Poisson) should give very similar conclusions.

When studying linear regression, our models were of the form

$$E[\mathbf{y}] = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_{p-1} \mathbf{x}_{p-1}.$$

The response $\mathbf{y}$ was continuous, not discrete, and we wanted to predict the mean response and explain the variability among the observed outcomes.

# Lecture 19: Random Effects Model

*Reading Assignment:*

- Muller and Fetterman, Chapter 15: "Special Cases of Two-Way ANOVA and Random Effects" (Required)
- Littell, Milliken, Stroup, and Wolfinger. *SAS System for Mixed Models*. (Great SAS reference for longitudinal data analysis.)

Mixed effects models are an extension of the GLM for correlated data. Applications include

- clustering by center, clinic, or primary care physician,
- responses measured on individual mice in a litter,
- meta-analysis to combine the results of many different studies, and

- repeated measures over time.

In addition, factors in an experiment may be considered *fixed* or *random*. We used *fixed effects* when we wish to make inferences only about the particular $k$ levels of the factor observed in the study. Examples include gender, treatment, or dose. We use *random effects* when these levels are a sample from a population of levels. Examples include subjects, observers, and clinic sites.

In order to decide whether a given factor should be treated as random or fixed, consider repetition of the experiment. If you repeated an experiment testing a new drug on patients in 5 Chapel Hill psychiatric clinics, what would have to be the same to make the experiment a repeat? Presumably, the second experiment should have the same drugs and doses as the first (fixed factors), but using 5 different psychiatric clinics may not really matter (random factor).

So far, we have considered models of the form

$$y = \text{fixed effects} + \text{error}.$$

The variance in $y$ is partitioned into what can be explained by fixed effects and what remains unexplained. The error term is the only part of the right hand side of that equation that has any variance. In addition, recall that we assume for the glm that error terms are independent.

If a model contains a random factor, then we have the form

$$y = \text{fixed effects} + \text{random effects} + \text{error}.$$

The random factor also has variance, and we now partition variance into that due to fixed factors, random factors, and unexplained factors. While we assume that subjects with different levels of the random factor are independent, we allow subjects at the same level of the random factor to have correlated errors.

## Introduction to Clustered Data

Multiple responses ($\mathbf{y}$'s), called *repeated measures*, are often taken per subject. If the repeated measures are recorded over time, we call these *longitudinal data*. The set of one *subject*'s repeated measures make up a *cluster*.

*Clustered data* do not have to be longitudinal. For example, clusters could be defined by members families participating in a genetic study or patients in medical clinics across North Carolina. In the family study, we assume subjects from different families will be independent, but that subjects within a family may be correlated.

Clustered data generally signal a violation of the homogeneity of variances assumption because subjects within a cluster are typically more alike than subjects in different clusters. Regression procedures must take into account the correlation between subjects within a cluster, and assuming falsely that observations in a cluster are

independent may give invalid results.

The correlation (dependence) structure in the data takes a specific form. Clustered data imply that observations inside a cluster (family, subject) are correlated but observations from different clusters are uncorrelated.

# Longitudinal Analysis

*Longitudinal* studies have designs in which the outcome variable is measured repeatedly over time.

Examples include

- The ARIC (Atherosclerosis Risk in Communities) Study at CSCC, in which over 15,000 subjects were examined at baseline and every three years for a number of cardiovascular endpoints.
- A study at the CPC that monitors women during pregnancy and for one year after delivery to investigate factors related to lifetime weight gain.
- HIV clinical trials in which viral load levels are monitored throughout the course of a patient's disease.

Scientific advantages of longitudinal study designs include the following.

- Longitudinal studies allow investigation of events that occur over time, which is essential to the study of growth, aging, or the course of disease.
- Longitudinal studies allow us to study the order of events.
- Longitudinal studies permit more complete ascertainment of exposure histories in epidemiologic studies.
- Longitudinal studies can reduce unexplained variability in response by using the subject as his or her own control (crossover studies).

## Randomized-blocks Experiment

*Blocking* is a powerful experimental design technique when the experimental subjects are heterogeneous with respect to certain variables that are associated with the response but are not of primary interest. Blocking consists of the following two steps:

1. grouping homogeneous experimental units together to form a block, and
2. assigning treatments at random to experimental units within a block.

Blocking prevents one type of subject from predominantly receiving one type of treatment. Suppose we conduct a national clinical trial of three treatments in 100 hospitals across the United States. Because we feel that patients within a hospital may be more alike than patients across hospitals (due to factors like population served by hospital, quality of nursing care in hospital, etc.), we wish to treat each hospital

as a block. Within each hospital (block), we will randomly assign the three treatments.

## Components of Variance

Until now, we have considered only one variance, $\sigma^2$, in our regression models. When we use random effects, we have additional variance terms to estimate. For example, we may treat hospitals as random factors in a multi-center study if we view these hospitals as random samples from a population of US hospitals (but aren't partically interested in hospital performance). If the variance due to the hospitals is much larger than the random error, for example, then this indicates that hospitals are extremely variable, and future studies would benefit by choosing a greater number of hospitals. If hospital variance is low, then we might have saved money by recruiting more patients in fewer centers.

## General Linear Mixed-Effects Model

The *general linear mixed-effects model* is often written

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i,$$

where $\mathbf{Y}_i$ is the vector of responses for subject $i$, $\mathbf{X}_i$ is the vector of fixed covariates for subject $i$, $\boldsymbol{\beta}$ is a p-dimensional parameter vector, $\mathbf{Z}_i$ is a matrix of known covariates, and $\mathbf{b}_i$ is a $q$-dimensional vector containing the random effects. We assume that $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D})$, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{i})$, and $\mathbf{b}_1, \ldots, \mathbf{b}_n, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are independent. This model is also commonly called the Laird-Ware model after its developers (Laird and Ware, 1982).

In BIOS 663, we will just consider $\mathbf{Z}_i$ to be indicator variables denoting cluster (e.g., family) membership, though more complicated structures are possible.

## Randomized Block Design Example

Three drugs that were thought to affect lymphocyte production were included in an experiment along with a placebo drug (a control that has no physiological properties). To control for potential variation among experimental units (here, mice), it was decided to block together mice from the same litter (same mother and father and born at the same time). Five (5) litters of mice were used where four (4) mice were selected from each litter, then one of the four drugs was randomly assigned (without replacement) to each mouse within each litter. This resulted in a balanced randomized complete block design. A complete block is a block that receives all levels of the treatments. After a sufficient period of time, a blood sample was drawn from each mouse and the number of lymphocytes per cubic millimeter was determined. The data are represented as thousands of lymphocytes per cubic millimeter. The data are taken from Mead et al. (Mead, R., R.N. Curnow, and A.M. Hasted. 1993. Statistical Methods in Agriculture

and Experimental Biology. Chapman and Hall, London, p656).

Note that the randomization of assignments of drug levels to mice is performed independently for each litter. Since the drugs that are used in the experiment were purposefully selected and interest will be in the mean response of each drug it is reasonable to treat the drugs as fixed effects. The litters of mice, however, should be considered as a random effect. If the experiment were repeated again it is unlikely that these same litters would be used again. Further, interest would be in making inferences over all potential litters of mice rather than just the 5 litters used in this experiment. This randomized complete block model could be written as:

$$y_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij},$$

where $i = 1, \ldots, 5$ and $j = 1, \ldots, 4$. As usual, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and we let $b_i \sim N(0, \sigma_b^2)$. $y_{ij}$ is the lymphocyte count on the mouse in litter $i$ receiving drug $j$, $\mu$ is the overall mean, $\alpha_j$ is the effect of the jth drug, $b_i$ is the effect of the ith block (i.e., the ith litter), and $b_i$ and $e_{ij}$ are

independent random variables. Note that we have assumed that the block effect and the treatment effect do not interact. It is the investigators responsibility to select blocks appropriate for the experiment that will not interact with the treatment. If interaction is possible, then it may be better to include the "block" in the experiment as a random treatment factor and use a factorial design so that the interaction can be investigated.

As there is a control treatment in the experiment, it may be desired to make comparisons of each treatment (drug) mean with the control level to determine separately for each drug whether or not it had an effect. This might be important in a drug screening study where interest is in determining which drugs to continue to pursue, rather than comparing among drugs to determine which is best or which ones differ in thier effects. The Dunnetts multiple comparison procedure is useful in this instance. Dunnetts procedure controls the type I error rate over the family of statements of each mean to the control. Note that comparisons of this type can also be oneided comparisons as you

may only be interested in differences of means versus the control that are positive (or negative) (drug mean is larger than the control, drug mean is less than the control). PROC MIXED permits both two and one sided comparisons. For two sided comparisons the LSMEANS statement is coded as

```
LSMEANS effect / DUNNETTS PDIFF=CONTROL('control level value');
```

while the oneided comparisons are coded as

```
LSMEANS effect / DUNNETTS PDIFF=CONTROLU('control level value');
```

where treatments are expected to be higher than the control, or

```
LSMEANS effect / DUNNETTS PDIFF=CONTROLL('control level value');
```

where treatments are expected to be below the control. In this experiment we will assume that we want the drugs to enhance (increase) the production of lymphocytes so we will use the oneided, upper comparison coding. Of course, these decisions should be made

prior to performing the data analysis and should be written in as a part of the project proposal.

Before we proceed with the analysis, let's investigate some of the ways in which blocking will affect the results and our summaries of the data, and hopefully we will also gain some insight into how blocking works and when to decide to use it. First, let's compute the variance of an individual observation $y_{ij}$
$Var(y_{ij}) = Var(\mu + \alpha_j + b_i + e_{ij}) = Var(b_i + e_{ij}) =$
$Var(b_i) + Var(e_{ij}) + Cov(b_i, e_{ij}) = \sigma_b^2 + \sigma^2$.
From this result we can compute the variance of the means of the drugs as
$Var(\bar{y}_j) = \frac{1}{n^2} \sum_i Var(y_{ij}) = \frac{1}{n^2} \sum_i (\sigma_b^2 + \sigma^2) = \frac{1}{n}(\sigma_b^2 + \sigma^2)$
where n = 5 is the number of blocks or litters in the experiment. Now let's investigate the covariance and correlation between measurements. Note that measurements made on mice from different litters are independent as defined by our model since both random effects are independently distributed. However, mice from the same litter are not

independent as they share the common block or litter random effect. Thus the covariance between mice on treatments (drugs) $j$ and $j'$ from litter $i$ is given by:

$Cov(y_{ij}, y_{ij'}) = Cov(\mu + \alpha_j + b_i + e_{ij}, \mu + \alpha'_j + b_i + e_{ij'})$
$= Cov(b_i + e_{ij}, b_i + e_{ij'}) = Var(b_j) = \sigma_b^2.$

```
*****************************************************************
* lympho.sas *;
* *;
* To study the effects of four drugs (A, B, C, and D, where D is a
* placebo), four (4) mice from each of five (5) litters were used,
* where litters were treated as blocks. Drugs were randomly *;
* assigned to the mice within each litter. Lymphocyte counts *;
* (thousands per cubic millimeter) were measured on each mouse. *;
* *;
* Source: Mead, R., R.N. Curnow, and A.M. Hasted. 1993. Statistical
* Methods in Agriculture and Experimental Biology. Chapman *;
```

```
* and Hall, London, p656. S540.S7.M4.1993 *;
*****************************************************************
Title1 Drug Effects on Lymphocyte Counts;
options ps=55 ls=80 pageno=1 nodate;
data lympho;
input drug $ litter lcytes;
label drug=drug litter=litter lcytes=lymphocyte counts;
datalines;
A 1 7.1
B 1 6.7
C 1 7.1
D 1 6.7
A 2 6.1
B 2 5.1
C 2 5.8
D 2 5.4
A 3 6.9
```

```
B 3 5.9
C 3 6.2
D 3 5.7
A 4 5.6
B 4 5.1
C 4 5.0
D 4 5.2
A 5 6.4
B 5 5.8
C 5 6.2
D 5 5.3
;

proc print data=lympho label;
run;

proc plot data=lympho;
```

```
plot lcytes*litter=* $ drug;
run;

/* Fit an RBD Model */
proc mixed data=lympho method=REML covtest;
class litter drug;
model lcytes = drug;
random litter;
lsmeans drug / adjust=dunnett pdiff=control(D);
/* compare drugs to the placebo level, oneailed test */
lsmeans drug / adjust=dunnett pdiff=controlu(D);
run;
```

```
Drug Effects on Lymphocyte Counts                          20
The Mixed Procedure
Model Information
Data Set                     WORK.LYMPHO
Dependent Variable           LCytes
```

```
Covariance Structure          Variance Components
Estimation Method             REML
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Containment

Class Level Information
Class      Levels    Values
Litter          5    1 2 3 4 5
Drug            4    A B C D


Dimensions
Covariance Parameters              2
Columns in X                       5
Columns in Z                       5
Subjects                           1
```

```
Max Obs Per Subject                20

Number of Observations

Number of Observations Read                20

Number of Observations Used                20

Number of Observations Not Used             0


Iteration History

Iteration    Evaluations    -2 Res Log Like       Criterion

0              1        38.70809588

1              1        18.49496651      0.00000000

Convergence criteria met.
```

Drug Effects on Lymphocyte Counts 21

The Mixed Procedure

Covariance Parameter Estimates

| Cov Parm | Estimate | Standard Error | Z Value | Pr Z |
|---|---|---|---|---|
| Litter | 0.3869 | 0.2830 | 1.37 | 0.0858 |
| Residual | 0.05308 | 0.02167 | 2.45 | 0.0072 |

Fit Statistics

| | |
|---|---|
| -2 Res Log Likelihood | 18.5 |
| AIC (smaller is better) | 22.5 |
| AICC (smaller is better) | 23.4 |
| BIC (smaller is better) | 21.7 |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|

| | | | | |
|---|---|---|---|---|
| Drug | 3 | 12 | 11.59 | 0.0007 |

Least Squares Means

| Effect | Drug | Estimate | Standard Error | DF | t Value | Pr > \| |
|---|---|---|---|---|---|---|
| Drug | A | 6.4200 | 0.2966 | 12 | 21.64 | <.00 |
| Drug | B | 5.7200 | 0.2966 | 12 | 19.28 | <.00 |
| Drug | C | 6.0600 | 0.2966 | 12 | 20.43 | <.00 |
| Drug | D | 5.6600 | 0.2966 | 12 | 19.08 | <.00 |
| Drug | A | 6.4200 | 0.2966 | 12 | 21.64 | <.00 |
| Drug | B | 5.7200 | 0.2966 | 12 | 19.28 | <.00 |
| Drug | C | 6.0600 | 0.2966 | 12 | 20.43 | <.00 |
| Drug | D | 5.6600 | 0.2966 | 12 | 19.08 | <.00 |

Differences of Least Squares Means

Standard

| Effect | Drug | Drug | Estimate | Error | DF | t Value | Tails |
|---|---|---|---|---|---|---|---|
| Drug | A | D | 0.7600 | 0.1457 | 12 | 5.22 | Both |
| Drug | B | D | 0.06000 | 0.1457 | 12 | 0.41 | Both |
| Drug | C | D | 0.4000 | 0.1457 | 12 | 2.75 | Both |
| Drug | A | D | 0.7600 | 0.1457 | 12 | 5.22 | Upper |
| Drug | B | D | 0.06000 | 0.1457 | 12 | 0.41 | Upper |
| Drug | C | D | 0.4000 | 0.1457 | 12 | 2.75 | Upper |

Differences of Least Squares Means

| Effect | Drug | Drug | Adjustment | Adj P |
|---|---|---|---|---|
| Drug | A | D | Dunnett-Hsu | 0.0006 |
| Drug | B | D | Dunnett-Hsu | 0.9543 |
| Drug | C | D | Dunnett-Hsu | 0.0448 |
| Drug | A | D | Dunnett-Hsu | 0.0003 |
| Drug | B | D | Dunnett-Hsu | 0.5844 |
| Drug | C | D | Dunnett-Hsu | 0.0224 |

## "Screamer" Study

A study was designed to investigate different strategies for reducing disruptive vocalizations among nursing home residents with Alzheimer's disease. Each resident in the study received four treatments (3 interventions and a control). This *crossover study* (patients received all the treatments) was designed with a "washout" period between treatments so that there would be no "carryover" effects.

Study subjects were nursing home residents, and the observational units were resident responses on each treatment. There were twelve residents (clusters) aged 71-87, 11 of whom were female. We have 4 observations (treatment responses) in each cluster. We call treatment the *crossover factor*.

This study design, in which each subject serves as his/her own control, is more powerful than a study that randomizes one-fourth (or 3)

residents to one of the four treatments. The ability to make within-subject comparisons allows us to control for much person-to-person variability.

The response of interest is the percentage of 120 fifteen-second audiotape samples with vocalizations above a fixed decibel level. The four treatments (administered on different days) were a control, a stuffed teddy bear, headphones with music, and both the teddy bear and headphones with music.

We wish to know if there are any differences in the effectiveness of the treatments.

Consider the following hypothetical subset of the data.

| Subject | Control | Bear | Music | Both |
|---|---|---|---|---|
| 1 | 45 | 33 | 8 | 37 |
| 5 | 23 | 16 | 13 | 9 |
| 6 | 22 | 28 | 31 | 12 |
| 7 | 46 | 46 | 30 | 16 |
| 8 | 68 | 58 | 68 | 48 |
| 10 | 13 | 13 | 7 | 8 |
| 11 | 18 | 27 | 17 | 9 |
| 13 | 35 | 32 | 51 | 4 |
| 15 | 49 | 44 | 21 | 27 |
| 16 | 57 | 57 | 74 | 16 |
| 19 | 36 | 30 | 46 | 30 |
| 21 | 33 | 21 | 21 | 21 |

Using PROC MEANS in SAS, we obtain the following summary statistics.

| Treatment | Mean | Std Dev |
|---|---|---|
| Control | 37.08 | 16.59 |
| Bear | 33.75 | 14.73 |
| Music | 32.25 | 22.66 |
| Both | 19.75 | 13.35 |

We will again use the model

$$y_{ij} = \mu + \alpha_j + b_i + \varepsilon_{ij},$$

where $i = 1, \ldots, 12$ and $j = 1, \ldots, 4$. As usual, $\varepsilon_{ij} \sim N(0, \sigma^2)$. But how about the distribution of $b_i$?

## Correlation

Two variables, $Y_1$ and $Y_2$, are *independent* if the conditional distribution of $Y_1 \mid Y_2$ does not depend on $Y_2$.

Two variables, $Y_1$ and $Y_2$, are *uncorrelated* (i.e.,not *linearly* dependent) if $E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})] = 0$. Note that variables can be uncorrelated without being independent (e.g.,$Y_1 = (-1, 0, 1)$ and $Y_2 = (1, 0, 1)$).

Two variables are *correlated* if $E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})] \neq 0$. The *covariance* between $Y_1$ and $Y_2$ is given by $E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})]$. Covariance can take any value and will be dependent on the units of the variables of interest. To make it independent of the units, we can divide by the standard deviations of the two variables to obtain the

correlation, which must lie between -1 and 1:

$$\text{Corr}(Y_1, Y_2) = \frac{E[(Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2})]}{\sigma_{Y_1}\sigma_{Y_2}}.$$

Repeated measures from the same person are usually *positively* correlated.

We use SAS to calculate the covariance and correlation matrices for the screamer study below.

```
data new;
input subject control bear music both;
cards;
1    45    33    8    37
5    23    16    13    9
6    22    28    31    12
7    46    46    30    16
8    68    58    68    48
10    13    13    7    8
11    18    27    17    9
13    35    32    51    4
```

```
15   49    44    21    27
16   57    57    74    16
19   36    30    46    30
21   33    21    21    21
;


proc corr data=new cov;
var control bear music both;
run;
*****************************************************************************

                        The CORR Procedure

            4  Variables:    control  bear     music    both

                      Covariance Matrix, DF = 11

                     control      bear       music       both
      control        275.3        223.9      251.9       160.1
      bear           223.9        217.1      249.7       103.4
      music          251.9        249.7      513.6       79.1
```

| | | | | |
|---|---|---|---|---|
| both | 160.1 | 103.4 | 79.1 | 178.2 |

Simple Statistics

| Variable | N | Mean | Std Dev | Sum | Minimum |
|---|---|---|---|---|---|
| control | 12 | 37.08333 | 16.59386 | 445.00000 | 13.00000 |
| bear | 12 | 33.75000 | 14.73478 | 405.00000 | 13.00000 |
| music | 12 | 32.25000 | 22.66405 | 387.00000 | 7.00000 |
| both | 12 | 19.75000 | 13.34933 | 237.00000 | 4.00000 |

Pearson Correlation Coefficients, N = 12
Prob > |r| under H0: Rho=0

| | control | bear | music |
|---|---|---|---|
| control | 1.00000 | 0.91585 | 0.67000 |
| | | <.0001 | 0.0171 |
| bear | 0.91585 | 1.00000 | 0.74800 |
| | <.0001 | | 0.0051 |
| music | 0.67000 | 0.74800 | 1.00000 |
| | 0.0171 | 0.0051 | |
| both | 0.72281 | 0.52607 | 0.26164 |
| | 0.0079 | 0.0789 | 0.4114 |

We notice that the independence assumption is not valid.

Given the responses for subject $i$, we define the *covariance matrix* as the following symmetric matrix of variances and covariances:

$$\text{Cov}\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & & \sigma_{2k} \\ \vdots & & \vdots & \\ \sigma_{1k} & \sigma_{2k} & \dots & \sigma_{kk} \end{bmatrix}.$$

Note that this matrix contains $\frac{k(k+1)}{2}$ unique elements. We sometimes refer to this as an *unstructured* covariance matrix. Sometimes it is advantageous to model the covariance structure more parsimoniously (especially when there are many repeated measurements). A popular covariance matrix that allows correlation to deteriorate over time is an *autoregressive* covariance matrix. For example, the first-order autoregressive or AR(1) covariance matrix is given by

$$\text{Cov}\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ y_{ik} \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{k-1} \\ \rho & 1 & \rho & & \rho^{k-2} \\ \vdots & & & \vdots & \\ \rho^{k-1} & \rho^{k-2} & \dots & \rho & 1 \end{bmatrix}$$

and contains 2 parameters.

Perhaps the most popular choice for mixed-model analysis of variance assumes that the correlation between repeated measurements is due to each subject's underlying level of response. This *subject effect* is treated as a random variable in mixed-model ANOVA. For example, if the expected response to treatment is given by $E(Y_{ij}) = \mathbf{X}_{ij}\boldsymbol{\beta}$, then the response for subject $i$ differs from the population mean by a subject effect, $b_i$, and a within-subject error, $w_{ij}$, leading to the model

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + b_i + w_{ij},$$

where $b_i \sim N(0, \sigma_b^2)$ and $w_{ij} \sim N(0, \sigma_w^2)$. The covariance matrix for this model has the *compound symmetry* form:

$$\text{Cov}\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ik} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 + \sigma_w^2 & \sigma_b^2 & \dots \sigma_b^2 & \\ \sigma_b^2 & \sigma_b^2 + \sigma_w^2 & & \sigma_b^2 \\ \vdots & & \vdots & \\ \sigma_b^2 & \sigma_b^2 & \dots & \sigma_b^2 + \sigma_w^2 \end{bmatrix}$$

and contains only 2 parameters. Although this is not valid under as wide a set of circumstances as the unrestricted covariance, at times the unrestricted covariance can contain too many parameters to estimate from the data.

It can be shown that the correlation between two subjects in the same group under the CS structure is given by $\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$. This is known as an exchangeable correlation. In this case, the $\alpha$'s are fixed treatment

effects, and the $b$'s represent each subject's own response tendency (some people may be more vocal than others, and thus their responses may tend to be high regardless of "treatment").

One option for our analysis is to use SAS PROC MIXED to extend PROC GLM and allow for our clusters of correlated observations. Here, we define the *cluster* to be the study subject. Responses across clusters, or of different subjects, are assumed to be independent. Responses within a cluster, or on an individual subject, are assumed to be correlated.

The following SAS code was used to input the data and fit the mixed effects model to control for correlation among members in a cluster.

```
data new2(keep=subject trt scream);
set new;
trt=control; scream=control; output;
trt=bear; scream=bear; output;
trt=music; scream=music; output;
trt=both; scream=both; output;
run;
```

```
proc mixed data=new2;
class subject trt;
model scream=trt/s;
repeated/type=cs subject=subject r;
contrast bear-both trt 1 -1 0 0;
contrast bear-control trt 1 0 -1 0;
contrast bear-music trt 1 0 0 -1;
contrast both-control trt 0 1 -1 0;
contrast both-music trt 0 1 0 -1;
contrast control-music trt 0 0 1 -1;
run;
```

The output is provided below.

```
                    The Mixed Procedure

                     Model Information

        Data Set                     WORK.NEW2
        Dependent Variable           scream
```

```
Covariance Structure          Compound Symmetry
Subject Effect                subject
Estimation Method             REML
Residual Variance Method      Profile
Fixed Effects SE Method       Model-Based
Degrees of Freedom Method     Between-Within


                Class Level Information

 Class      Levels    Values

 subject        12    1 5 6 7 8 10 11 13 15 16 19 21
 trt             4    bear both control music


                     Dimensions

          Covariance Parameters            2
          Columns in X                     5
          Columns in Z                     0
          Subjects                        12
```

| | |
|---|---|
| Max Obs Per Subject | 4 |
| Observations Used | 48 |
| Observations Not Used | 0 |
| Total Observations | 48 |

Iteration History

| Iteration | Evaluations | -2 Res Log Like | Criterion |
|---|---|---|---|
| 0 | 1 | 385.19441916 | |
| 1 | 1 | 366.18072711 | 0.00000000 |

Convergence criteria met.

Estimated R Matrix for subject 1

| Row | Col1 | Col2 | Col3 | Col4 |
|---|---|---|---|---|
| 1 | 296.08 | 178.08 | 178.08 | 178.08 |

| | | | | |
|---|---|---|---|---|
| 2 | 178.08 | 296.08 | 178.08 | 178.08 |
| 3 | 178.08 | 178.08 | 296.08 | 178.08 |
| 4 | 178.08 | 178.08 | 178.08 | 296.08 |

The Mixed Procedure

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate |
|---|---|---|
| CS | subject | 178.08 |
| Residual | | 118.01 |

Fit Statistics

| | |
|---|---|
| -2 Res Log Likelihood | 366.2 |
| AIC (smaller is better) | 370.2 |
| AICC (smaller is better) | 370.5 |
| BIC (smaller is better) | 371.2 |

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 1 | 19.01 | <.0001 |

Solution for Fixed Effects

| Effect | trt | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | | 32.2500 | 4.9673 | 11 | 6.49 | <.0001 |
| trt | bear | 1.5000 | 4.4349 | 33 | 0.34 | 0.7373 |
| trt | both | -12.5000 | 4.4349 | 33 | -2.82 | 0.0081 |
| trt | control | 4.8333 | 4.4349 | 33 | 1.09 | 0.2837 |
| trt | music | 0 | . | . | . | . |

Type 3 Tests of Fixed Effects

| | Num | Den |
|---|---|---|

| Effect | DF | DF | F Value | Pr > F |
|---|---|---|---|---|
| trt | 3 | 33 | 5.84 | 0.0026 |

Contrasts

| Label | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| bear-both | 1 | 33 | 9.97 | 0.0034 |
| bear-control | 1 | 33 | 0.56 | 0.4576 |
| bear-music | 1 | 33 | 0.11 | 0.7373 |
| both-control | 1 | 33 | 15.28 | 0.0004 |
| both-music | 1 | 33 | 7.94 | 0.0081 |
| control-music | 1 | 33 | 1.19 | 0.2837 |

We see that the treatments are indeed different. Specifically, the teddy bear and music together are significantly better than all the other treatments.

SAS code and output for PROC MIXED with the UN structure are given below.

```
proc mixed data=new2 dfbw noclprint;
class subject trt;
model sbp=trt/s;
repeated/type=un subject=subject r;
contrast bear-both trt 1 -1 0 0;
contrast bear-control trt 1 0 -1 0;
contrast bear-music trt 1 0 0 -1;
contrast both-control trt 0 1 -1 0;
contrast both-music trt 0 1 0 -1;
contrast control-music trt 0 0 1 -1;
run;
*****************************************************************
```

```
                    The Mixed Procedure

                     Model Information
                Data Set                     WORK.NEW2
                Dependent Variable           scream
                Covariance Structure         Unstructured
```

| | |
|---|---|
| Subject Effect | subject |
| Estimation Method | REML |
| Residual Variance Method | None |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|---|---|
| Covariance Parameters | 10 |
| Columns in X | 5 |
| Columns in Z | 0 |
| Subjects | 12 |
| Max Obs Per Subject | 4 |

Number of Observations

| | |
|---|---|
| Number of Observations Read | 48 |
| Number of Observations Used | 48 |
| Number of Observations Not Used | 0 |

Iteration History

| Iteration | Evaluations | -2 Res Log Like | Criterion |
|---|---|---|---|
| 0 | 1 | 385.19441916 | |
| 1 | 1 | 340.32444257 | 0.00000000 |

Convergence criteria met.

Estimated R Matrix for subject 1

| Row | Col1 | Col2 | Col3 | Col4 |
|---|---|---|---|---|
| 1 | 275.36 | 223.93 | 251.98 | 160.11 |
| 2 | 223.93 | 217.11 | 249.80 | 103.48 |
| 3 | 251.98 | 249.80 | 513.66 | 79.1591 |
| 4 | 160.11 | 103.48 | 79.1591 | 178.20 |

The Mixed Procedure

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate |
|---|---|---|
| UN(1,1) | subject | 275.36 |
| UN(2,1) | subject | 223.93 |

| | | |
|---|---|---|
| UN(2,2) | subject | 217.11 |
| UN(3,1) | subject | 251.98 |
| UN(3,2) | subject | 249.80 |
| UN(3,3) | subject | 513.66 |
| UN(4,1) | subject | 160.11 |
| UN(4,2) | subject | 103.48 |
| UN(4,3) | subject | 79.1591 |
| UN(4,4) | subject | 178.20 |

Fit Statistics

| | |
|---|---|
| -2 Res Log Likelihood | 340.3 |
| AIC (smaller is better) | 360.3 |
| AICC (smaller is better) | 367.0 |
| BIC (smaller is better) | 365.2 |

```
  Null Model Likelihood Ratio Test
DF    Chi-Square      Pr > ChiSq
9         44.87          <.0001
```

Solution for Fixed Effects

| Effect | trt | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | | 32.25 | 6.54 | 11 | 4.93 | 0.0005 |
| trt | bear | 1.50 | 4.38 | 11 | 0.34 | 0.7390 |
| trt | both | -12.50 | 6.66 | 11 | -1.87 | 0.0876 |
| trt | control | 4.83 | 4.87 | 11 | 0.99 | 0.3427 |
| trt | music | 0 | . | . | . | . |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| trt | 3 | 11 | 11.70 | 0.0010 |

The Mixed Procedure

Contrasts

| Label | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| bear-both | 1 | 11 | 12.49 | 0.0047 |
| bear-control | 1 | 11 | 2.99 | 0.1118 |
| bear-music | 1 | 11 | 0.12 | 0.7390 |
| both-control | 1 | 11 | 27.04 | 0.0003 |
| both-music | 1 | 11 | 3.51 | 0.0876 |
| control-music | 1 | 11 | 0.98 | 0.3427 |

We can check whether the compound symmetry assumption is valid for the screamer data by comparing the residual log likelihoods from the compound symmetry and unrestricted covariance models.

To test the adequacy of the compound symmetry assumption, we consider a chi-squared test of the difference between the two residual log likelihoods. Recall that the compound symmetric covariance estimates 2 parameters, while the unstructured covariance estimates $\frac{4(5)}{2} = 10$ parameters.

| | UN | CS |
|---|---|---|
| -2 Res Log L | 340.3 | 366.2 |
| df | 1 | 9 |

Thus we compare -2 log likelihood ratio=(366.2-340.3)=25.9 to the chi-squared distribution with 8 df (which has critical value 15.5) and therefore reject the null hypothesis that the smaller model (compound

symmetry) is valid. We conclude that the compound symmetry assumption is not valid for the data.

## Parallel Groups Repeated Measures Design

In the parallel groups design, two or more groups of subjects are measured repeatedly over time. For example, groups may be defined by *randomization* to one of several treatments.

We consider designs in which all subjects are meant to be measured at the same set of follow-up times, but these methods will allow for certain types of missing data.

The analysis goal is to characterize the patterns of change over time in the groups and to determine whether the patterns differ in the groups.

**Example: Schizophrenia Clinical Trial**

A multi-center study was conducted to assess the effect of two treatments and a placebo on decreasing the PANSS (positive and negative symptom scale) score. Patients were randomized to receive only one of the two treatments (haloperidol and risperidone) or a placebo. Measurements of the PANSS were taken at baseline and after 1, 2, 4, 6, and 8 weeks of treatment. Consider the following plot of PANSS score means at each measurement time. What do you think about the patterns of change over time?

PANSS Scores
Placebo
Haloperidol
Risperidone
Mean PANSS Score
100
95
90
85
80
75
70
0
2
4
6
8
Measurement Time

Notice that since this is a randomized study, the group means should be the same at baseline, so any test of treatment effect is essentially a test of treatment by time interaction (we cannot have parallel lines if there really is a treatment effect, since parallel lines would imply coincident lines in this case).

In the analysis of parallel groups data, we may have one or more of the following hypotheses.

- $H_0$: Are the profiles of the means similar in the three groups? (That is, are the line segments between adjacent measurement occasions parallel?) This hypothesis is the hypothesis of no group by time interaction.

- $H_0$: If the profiles are parallel, are they also at the same level? This is the hypothesis of no group effect. (In a randomized study, the profiles should all start at the same level.)

- $H_0$: If the profiles are parallel, are the means constant over time? This is the hypothesis of no time effect.

The appropriate hypotheses for any given dataset must be derived from the relevant scientific issues in that study. For example, we are interested only in the

first and last hypotheses for this randomized study. (Why?)

## Notation for Parallel Groups Repeated Measures

Changing notation slightly, let $n$ be the number of subjects (clusters) and $N$ be the total number of observations. We consider the univariate representation (one row for each observation of the outcome) of the multivariate data.

If there are $k$ measurement occasions, we define $k-1$ indicator variables. For the $i^{th}$ observation, we let $x_{ij} = 1$ if the observation was taken at time $j$ and 0 otherwise, $j = 1, \ldots, k-1$. In addition, let $x_{ik}$ be the indicator variable for the haloperidol group and $x_{i,k+1}$ be the indicator variable for the placebo group. (Risperidone is the reference group for treatment.)

We also define $2(k-1)$ interaction terms as products of the time and group indicator variables. With these terms in the model, our model has $1 + (k-1) + (2) + 2(k-1) = 3k$ parameters for the mean model. If we use an unrestricted covariance matrix, we have $\frac{6(7)}{2} = 21$ additional parameters for the covariance.

If we let $i$ index all the responses, $i = 1, \ldots, N$, then the model is given by

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} \\ &\quad + \beta_6 x_{i6} + \beta_7 x_{i7} + \beta_8 x_{i8} + \beta_9 x_{i9} + \beta_{10} x_{i10} \\ &\quad + \beta_{11} x_{i11} + \beta_{12} x_{i12} + \beta_{13} x_{i13} + \beta_{14} x_{i14} \\ &\quad + \beta_{15} x_{i15} + \beta_{16} x_{i16} + \beta_{17} x_{i17} + \varepsilon_i, \end{aligned}$$

where

- $x_{i1}, \ldots x_{i5}$ represent measurement times 0, 1, 2, 4, and 6 (with time 8 as the reference),
- $x_{i6}$ and $x_{i7}$ are the indicator variables for haloperidol and placebo, respectively (risperidone is the reference),
- $x_{i8}, \ldots, x_{i12}$ are the interactions between haloperidol and time, and
- $x_{i13}, \ldots, x_{i17}$ are the interactions between placebo and time.

## Assumptions of the Parallel Groups Model

1. The subjects are random samples from each of the groups.
2. Observations from different individuals are independent, while repeated measurements on the same individual are not assumed to be independent.
3. The vector of outcomes for a given subject has a multivariate normal distribution.
4. The expected values of the individual observations are given by the linear regression model.
5. If observations are missing, they are missing at random (MAR).

**Example: Schizophrenia Data**

The following SAS code may be used to fit this model to the schizophrenia data (assuming the data are already in univariate format). Included are statements necessary to estimate differences in cell means for various time and treatment combinations. Note the ordering of the contrast coefficients. The first six coefficients represent haloperidol (TRT=1) at the six times (in numerical order). The next six coefficients represent placebo (TRT=2) at the six times, and the last six coefficients represent risperidone. The "e" option instructs SAS to print information about the contrast coefficients so that you can check to be sure you coded them properly (always a good idea!).

```
proc mixed data=new2 noclprint;
class id trtnew time t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
estimate time 1 vs. baseline haloperidol1 - diff for placebo
        time*trtnew -1 1 0 0 0 0 1 -1 0 0 0 0 0 0 0 0 0 0/e;
estimate time 1 vs. baseline placebo - diff for risperidone
        time*trtnew 0 0 0 0 0 0 1 -1 0 0 0 0 -1 1 0 0 0 0/e;
estimate time 2 vs. baseline haloperidol1 - diff for placebo
        time*trtnew -1 0 1 0 0 0 1 0 -1 0 0 0 0 0 0 0 0 0/e;
```

```
estimate time 2 vs. baseline placebo - diff for risperidone
        time*trtnew 0 0 0 0 0 0 1 0 -1 0 0 0 -1 0 1 0 0 0/e;
estimate time 4 vs. baseline haloperidol1 - diff for placebo
        time*trtnew -1 0 0 1 0 0 1 0 0 -1 0 0 0 0 0 0 0 0/e;
estimate time 4 vs. baseline placebo - diff for risperidone
        time*trtnew 0 0 0 0 0 0 1 0 0 -1 0 0 -1 0 0 1 0 0/e;
estimate time 6 vs. baseline haloperidol1 - diff for placebo
        time*trtnew -1 0 0 0 1 0 1 0 0 0 -1 0 0 0 0 0 0 0 0/e;
estimate time 6 vs. baseline placebo - diff for risperidone
        time*trtnew 0 0 0 0 0 0 1 0 0 0 -1 0 -1 0 0 0 1 0/e;
estimate time 8 vs. baseline haloperidol1 - diff for placebo
        time*trtnew -1 0 0 0 0 1 1 0 0 0 0 -1 0 0 0 0 0 0 0/e;
estimate time 8 vs. baseline placebo - diff for risperidone
        time*trtnew 0 0 0 0 0 0 1 0 0 0 0 -1 -1 0 0 0 0 1/ e;
run;
```

Selected SAS output is provided below.

```
The Mixed Procedure

                          Model Information

        Data Set                     WORK.NEW2
```

| | |
|---|---|
| Dependent Variable | y |
| Covariance Structure | Unstructured |
| Subject Effect | id |
| Estimation Method | REML |
| Residual Variance Method | None |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|---|---|
| Covariance Parameters | 21 |
| Columns in X | 28 |
| Columns in Z | 0 |
| Subjects | 523 |
| Max Obs Per Subject | 6 |
| Observations Used | 2468 |
| Observations Not Used | 670 |
| Total Observations | 3138 |

Iteration History

| Iteration | Evaluations | -2 Res Log Like | Criterion |
|---|---|---|---|
| 0 | 1 | 21961.16171710 | |
| 1 | 2 | 20151.85618890 | 0.00610058 |
| 2 | 1 | 20096.23405069 | 0.00167392 |
| 3 | 1 | 20080.65977798 | 0.00027233 |
| 4 | 1 | 20078.27196287 | 0.00001373 |
| 5 | 1 | 20078.16116305 | 0.00000004 |
| 6 | 1 | 20078.16081868 | 0.00000000 |

Convergence criteria met.

Estimated R Matrix for id 1

| Row | Col1 | Col2 |
|---|---|---|
| 1 | 356.09 | 238.69 |
| 2 | 238.69 | 507.39 |

The Mixed Procedure

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate |
|---|---|---|
| UN(1,1) | id | 356.09 |
| UN(2,1) | id | 250.92 |
| UN(2,2) | id | 429.79 |
| UN(3,1) | id | 238.69 |
| UN(3,2) | id | 382.01 |
| UN(3,3) | id | 507.39 |
| UN(4,1) | id | 229.26 |
| UN(4,2) | id | 361.56 |
| UN(4,3) | id | 451.91 |
| UN(4,4) | id | 571.64 |
| UN(5,1) | id | 216.73 |
| UN(5,2) | id | 340.27 |
| UN(5,3) | id | 432.90 |
| UN(5,4) | id | 515.10 |
| UN(5,5) | id | 630.37 |
| UN(6,1) | id | 193.48 |
| UN(6,2) | id | 317.84 |

```
UN(6,3)     id          396.08
UN(6,4)     id          484.38
UN(6,5)     id          562.16
UN(6,6)     id          625.87
```

Fit Statistics

| | |
|---|---|
| -2 Res Log Likelihood | 20078.2 |
| AIC (smaller is better) | 20120.2 |
| AICC (smaller is better) | 20120.5 |
| BIC (smaller is better) | 20209.6 |

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 20 | 1883.00 | <.0001 |

Solution for Fixed Effects

| Effect | trtnew | time | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| Intercept | | | 78.6367 | 1.5171 | 519 | 51.83 | <.0001 |
| trtnew | 1 | | 6.9346 | 3.5402 | 519 | 1.96 | 0.0507 |
| trtnew | 2 | | 17.0886 | 3.7912 | 519 | 4.51 | <.0001 |
| trtnew | 3 | | 0 | . | . | . | . |

The Mixed Procedure

Solution for Fixed Effects

| Effect | trtnew | time | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| time | | 0 | 13.8419 | 1.4886 | 519 | 9.30 | <.0001 |
| time | | 1 | 5.6702 | 1.3052 | 519 | 4.34 | <.0001 |
| time | | 2 | 1.8247 | 1.2063 | 519 | 1.51 | 0.1310 |
| time | | 4 | 0.9563 | 1.0284 | 519 | 0.93 | 0.3529 |
| time | | 6 | 0.1495 | 0.8046 | 519 | 0.19 | 0.8527 |
| time | | 8 | 0 | . | . | . | . |
| trtnew*time | 1 | 0 | -5.8040 | 3.4774 | 519 | -1.67 | 0.0957 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| trtnew*time | 1 | 1 | -2.1313 | 3.0885 | 519 | -0.69 | 0.4905 |
| trtnew*time | 1 | 2 | -0.09359 | 2.8842 | 519 | -0.03 | 0.9741 |
| trtnew*time | 1 | 4 | -0.2524 | 2.4959 | 519 | -0.10 | 0.9195 |
| trtnew*time | 1 | 6 | -0.2067 | 1.9499 | 519 | -0.11 | 0.9156 |
| trtnew*time | 1 | 8 | 0 | . | . | . | . |
| trtnew*time | 2 | 0 | -17.0786 | 3.7332 | 519 | -4.57 | <.0001 |
| trtnew*time | 2 | 1 | -10.6730 | 3.3797 | 519 | -3.16 | 0.0017 |
| trtnew*time | 2 | 2 | -6.6133 | 3.1757 | 519 | -2.08 | 0.0378 |
| trtnew*time | 2 | 4 | -3.9971 | 2.7871 | 519 | -1.43 | 0.1521 |
| trtnew*time | 2 | 6 | 2.4792 | 2.2450 | 519 | 1.10 | 0.2700 |
| trtnew*time | 2 | 8 | 0 | . | . | . | . |
| trtnew*time | 3 | 0 | 0 | . | . | . | . |
| trtnew*time | 3 | 1 | 0 | . | . | . | . |
| trtnew*time | 3 | 2 | 0 | . | . | . | . |
| trtnew*time | 3 | 4 | 0 | . | . | . | . |
| trtnew*time | 3 | 6 | 0 | . | . | . | . |
| trtnew*time | 3 | 8 | 0 | . | . | . | . |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| trtnew | 2 | 519 | 10.96 | <.0001 |
| time | 5 | 519 | 8.20 | <.0001 |
| trtnew*time | 10 | 519 | 3.68 | <.0001 |

Coefficients for time 1 vs. baseline
haloperidol1 - diff for placebo

| Effect | trtnew | time | Row1 |
|---|---|---|---|
| Intercept | | | |
| trtnew | 1 | | |
| trtnew | 2 | | |
| trtnew | 3 | | |
| time | | 0 | |

The Mixed Procedure

```
 Coefficients for time 1 vs. baseline
    haloperidol1 - diff for placebo

Effect          trtnew    time      Row1

time                      1
time                      2
time                      4
time                      6
time                      8
trtnew*time     1         0           -1
trtnew*time     1         1            1
trtnew*time     1         2
trtnew*time     1         4
trtnew*time     1         6
trtnew*time     1         8
trtnew*time     2         0            1
trtnew*time     2         1           -1
trtnew*time     2         2
trtnew*time     2         4
trtnew*time     2         6
trtnew*time     2         8
trtnew*time     3         0
```

```
trtnew*time    3         1
trtnew*time    3         2
trtnew*time    3         4
trtnew*time    3         6
trtnew*time    3         8


 Coefficients for time 1 vs. baseline
    placebo - diff for risperidone

Effect         trtnew    time      Row1

Intercept
trtnew         1
trtnew         2
trtnew         3
time                     0
time                     1
time                     2
time                     4
time                     6
time                     8
trtnew*time    1         0
```

```
trtnew*time    1         1
trtnew*time    1         2
trtnew*time    1         4
trtnew*time    1         6
trtnew*time    1         8

          The Mixed Procedure

 Coefficients for time 1 vs. baseline
    placebo - diff for risperidone

Effect         trtnew    time      Row1

trtnew*time    2         0            1
trtnew*time    2         1           -1
trtnew*time    2         2
trtnew*time    2         4
trtnew*time    2         6
trtnew*time    2         8
trtnew*time    3         0           -1
trtnew*time    3         1            1
trtnew*time    3         2
trtnew*time    3         4
```

```
                trtnew*time    3         6
                trtnew*time    3         8


.....some output omitted.....
```

Estimates

| Label | Estimate |
|---|---|
| time 1 vs. baseline haloperidol1 - diff for placebo | -2.7329 |
| time 1 vs. baseline placebo - diff for risperidone | -6.4056 |
| time 2 vs. baseline haloperidol1 - diff for placebo | -4.7549 |

Estimates

| Label | Standard Error | DF |
|---|---|---|
| time 1 vs. baseline haloperidol1 - diff for placebo | 2.5776 | 519 |
| time 1 vs. baseline placebo - diff for risperidone | 2.0356 | 519 |
| time 2 vs. baseline haloperidol1 - diff for placebo | 3.1184 | 519 |

Estimates

| Label | t Value | Pr > \|t\| |
|---|---|---|
| time 1 vs. baseline haloperidol1 - diff for placebo | -1.06 | 0.2895 |

```
time 1 vs. baseline placebo - diff for risperidone     -3.15    0.0017
time 2 vs. baseline haloperidol1 - diff for placebo    -1.52    0.1279
```

The Mixed Procedure

Estimates

| Label | Estimate |
|---|---|
| time 2 vs. baseline placebo - diff for risperidone | -10.4653 |
| time 4 vs. baseline haloperidol1 - diff for placebo | -7.5298 |
| time 4 vs. baseline placebo - diff for risperidone | -13.0814 |
| time 6 vs. baseline haloperidol1 - diff for placebo | -13.9604 |
| time 6 vs. baseline placebo - diff for risperidone | -19.5578 |
| time 8 vs. baseline haloperidol1 - diff for placebo | -11.2746 |
| time 8 vs. baseline placebo - diff for risperidone | -17.0786 |

Estimates

| Label | Standard Error | DF |
|---|---|---|

| | | |
|---|---|---|
| time 2 vs. baseline placebo - diff for risperidone | 2.4715 | 519 |
| time 4 vs. baseline haloperidol1 - diff for placebo | 3.5989 | 519 |
| time 4 vs. baseline placebo - diff for risperidone | 2.8744 | 519 |
| time 6 vs. baseline haloperidol1 - diff for placebo | 4.2460 | 519 |
| time 6 vs. baseline placebo - diff for risperidone | 3.3765 | 519 |
| time 8 vs. baseline haloperidol1 - diff for placebo | 4.6473 | 519 |
| time 8 vs. baseline placebo - diff for risperidone | 3.7332 | 519 |

Estimates

| Label | t Value | Pr > \|t\| |
|---|---|---|
| time 2 vs. baseline placebo - diff for risperidone | -4.23 | <.0001 |
| time 4 vs. baseline haloperidol1 - diff for placebo | -2.09 | 0.0369 |
| time 4 vs. baseline placebo - diff for risperidone | -4.55 | <.0001 |
| time 6 vs. baseline haloperidol1 - diff for placebo | -3.29 | 0.0011 |
| time 6 vs. baseline placebo - diff for risperidone | -5.79 | <.0001 |
| time 8 vs. baseline haloperidol1 - diff for placebo | -2.43 | 0.0156 |
| time 8 vs. baseline placebo - diff for risperidone | -4.57 | <.0001 |

What do we conclude based on the hypothesis tests given in the SAS output?

PANSS Scores
Placebo
Haloperidol
Risperidone
Mean PANSS Score
100
95
90
85
80
75
70
0
2
4
6
8
Measurement Time

Using the following SAS code and partial output, test whether the compound symmetry covariance is defensible for these data.

```
proc mixed data=new2 noclprint;
class id trtnew time t;
model y=trtnew time time*trtnew/s;
repeated t/type=cs subject=id r;
run;
*****************************************************************
                         The Mixed Procedure

                          Model Information

        Data Set                     WORK.NEW2
        Dependent Variable           y
        Covariance Structure         Compound Symmetry
        Subject Effect               id
        Estimation Method            REML
        Residual Variance Method     Profile
        Fixed Effects SE Method      Model-Based
        Degrees of Freedom Method    Between-Within
```

Dimensions

| | |
|---|---|
| Covariance Parameters | 2 |
| Columns in X | 28 |
| Columns in Z | 0 |
| Subjects | 523 |
| Max Obs Per Subject | 6 |
| Observations Used | 2468 |
| Observations Not Used | 670 |
| Total Observations | 3138 |

Iteration History

| Iteration | Evaluations | -2 Res Log Like | Criterion |
|---|---|---|---|
| 0 | 1 | 21961.16171710 | |
| 1 | 2 | 20548.58902724 | 0.00002967 |
| 2 | 1 | 20548.34199238 | 0.00000011 |
| 3 | 1 | 20548.34114692 | 0.00000000 |

Convergence criteria met.

```
Estimated R Matrix for id 1

Row        Col1        Col2

  1      455.62      302.14
  2      302.14      455.62


 Covariance Parameter Estimates

Cov Parm     Subject    Estimate

CS           id           302.14
Residual                  153.48

      The Mixed Procedure

        Fit Statistics

-2 Res Log Likelihood         20548.3
AIC (smaller is better)       20552.3
```

```
AICC (smaller is better)      20552.3
BIC (smaller is better)       20560.9
```

## Treating Time as a Continuous Variable

If the means tend to change linearly over time, we may wish to treat time as a continuous variable rather than as a categorical variable. If means do change linearly with time, then the treatment effect in a continuous-time model is captured in one single parameter, leading to more powerful tests.

If time is treated as continuous, the model is given by

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i,$$

where

- $X_{i1}$ represents measurement time as a continuous variable taking the values 0, 1, 2, 4, 6, or 8,
- $X_{i2}$ and $X_{i3}$ are the indicator variables for haloperidol and placebo, respectively,
- $X_{i4}$ is the interaction between haloperidol and time, and

- $X_{i5}$ is the interaction between placebo and time.

To fit this model, we remove TIME from the class statement and use the following SAS code.

```
proc mixed data=new2 noclprint;
class id trtnew t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
run;
```

Note that we still need the variable "t" (which takes the same values as "time") declared as a class variable so that we can tell SAS how the repeated measures are ordered within a subject. This is why we define two variables, "t" and "time," with the same values (if we wish to treat time as continuous, we still need a version of that variable that can be treated as a class variable).

For subjects on haloperidol, we have

$$E(Y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)TIME_i,$$

for subjects on placebo, we have

$$E(Y_i) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)TIME_i,$$

and for subjects on risperidone we have

$$E(Y_i) = \beta_0 + \beta_1 TIME_i.$$

Clearly, the model assumes each group's mean changes linearly over time.

## Selecting a Method for Parameterizing Time

Because fitting time as a class variable is equivalent to fitting a polynomial of order 5 in time, we can view the continuous time parameterization as a model nested in the categorical time parameterization. Thus a likelihood ratio test is appropriate for comparing the two models.

Using the following SAS code and selected output, test whether the continuous time model is appropriate for the schizophrenia data.

```
proc mixed data=new2 noclprint method=ml;
class id trtnew time t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
run;

proc mixed data=new2 noclprint method=ml;
class id trtnew t;
model y=trtnew time time*trtnew/s;
repeated t/type=un subject=id r;
run;
*****************************************************************
                        The Mixed Procedure
```

Model Information

| | |
|---|---|
| Data Set | WORK.NEW2 |
| Dependent Variable | y |
| Covariance Structure | Unstructured |
| Subject Effect | id |
| Estimation Method | ML |
| Residual Variance Method | None |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|---|---|
| Covariance Parameters | 21 |
| Columns in X | 28 |
| Columns in Z | 0 |
| Subjects | 523 |
| Max Obs Per Subject | 6 |
| Observations Used | 2468 |
| Observations Not Used | 670 |
| Total Observations | 3138 |

Iteration History

| Iteration | Evaluations | -2 Log Like | Criterion |
|---|---|---|---|
| 0 | 1 | 22020.87344360 | |
| 1 | 3 | 20159.01391755 | 0.00575612 |
| 2 | 2 | 20128.62705395 | 0.00065058 |
| 3 | 1 | 20123.02179920 | 0.00003529 |
| 4 | 1 | 20122.73459148 | 0.00000018 |
| 5 | 1 | 20122.73319055 | 0.00000000 |

Convergence criteria met.

Estimated R Matrix for id 1

| Row | Col1 | Col2 |
|---|---|---|
| 1 | 354.04 | 237.31 |
| 2 | 237.31 | 504.31 |

The Mixed Procedure

Covariance Parameter Estimates

| Cov Parm | Subject | Estimate |
|---|---|---|
| UN(1,1) | id | 354.04 |
| UN(2,1) | id | 249.48 |
| UN(2,2) | id | 427.29 |
| UN(3,1) | id | 237.31 |
| UN(3,2) | id | 379.79 |
| UN(3,3) | id | 504.31 |
| UN(4,1) | id | 227.95 |
| UN(4,2) | id | 359.46 |
| UN(4,3) | id | 449.18 |
| UN(4,4) | id | 567.96 |
| UN(5,1) | id | 215.48 |
| UN(5,2) | id | 338.29 |
| UN(5,3) | id | 430.28 |
| UN(5,4) | id | 511.81 |

| | | |
|---|---|---|
| UN(5,5) | id | 625.86 |
| UN(6,1) | id | 192.37 |
| UN(6,2) | id | 315.99 |
| UN(6,3) | id | 393.69 |
| UN(6,4) | id | 481.27 |
| UN(6,5) | id | 558.20 |
| UN(6,6) | id | 621.01 |

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 20122.7 |
| AIC (smaller is better) | 20200.7 |
| AICC (smaller is better) | 20202.0 |
| BIC (smaller is better) | 20366.9 |

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 20 | 1898.14 | <.0001 |

Solution for Fixed Effects

| Effect | trtnew | time | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| Intercept | | | 78.6367 | 1.5109 | 519 | 52.05 | <.0001 |
| trtnew | 1 | | 6.9346 | 3.5253 | 519 | 1.97 | 0.0497 |
| trtnew | 2 | | 17.0886 | 3.7749 | 519 | 4.53 | <.0001 |
| trtnew | 3 | | 0 | . | . | . | . |

The Mixed Procedure

Solution for Fixed Effects

| Effect | trtnew | time | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|
| time | | 0 | 13.8420 | 1.4824 | 519 | 9.34 | <.0001 |
| time | | 1 | 5.6703 | 1.2993 | 519 | 4.36 | <.0001 |
| time | | 2 | 1.8247 | 1.2006 | 519 | 1.52 | 0.1292 |
| time | | 4 | 0.9563 | 1.0231 | 519 | 0.93 | 0.3504 |
| time | | 6 | 0.1495 | 0.8002 | 519 | 0.19 | 0.8518 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| time | | 8 | 0 | . | . | . | . |
| trtnew*time | 1 | 0 | -5.8040 | 3.4627 | 519 | -1.68 | 0.0943 |
| trtnew*time | 1 | 1 | -2.1313 | 3.0743 | 519 | -0.69 | 0.4885 |
| trtnew*time | 1 | 2 | -0.09356 | 2.8704 | 519 | -0.03 | 0.9740 |
| trtnew*time | 1 | 4 | -0.2524 | 2.4830 | 519 | -0.10 | 0.9191 |
| trtnew*time | 1 | 6 | -0.2067 | 1.9392 | 519 | -0.11 | 0.9152 |
| trtnew*time | 1 | 8 | 0 | . | . | . | . |
| trtnew*time | 2 | 0 | -17.0786 | 3.7169 | 519 | -4.59 | <.0001 |
| trtnew*time | 2 | 1 | -10.6729 | 3.3639 | 519 | -3.17 | 0.0016 |
| trtnew*time | 2 | 2 | -6.6133 | 3.1602 | 519 | -2.09 | 0.0369 |
| trtnew*time | 2 | 4 | -3.9971 | 2.7726 | 519 | -1.44 | 0.1500 |
| trtnew*time | 2 | 6 | 2.4792 | 2.2327 | 519 | 1.11 | 0.2673 |
| trtnew*time | 2 | 8 | 0 | . | . | . | . |
| trtnew*time | 3 | 0 | 0 | . | . | . | . |
| trtnew*time | 3 | 1 | 0 | . | . | . | . |
| trtnew*time | 3 | 2 | 0 | . | . | . | . |
| trtnew*time | 3 | 4 | 0 | . | . | . | . |
| trtnew*time | 3 | 6 | 0 | . | . | . | . |
| trtnew*time | 3 | 8 | 0 | . | . | . | . |

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| trtnew | 2 | 519 | 11.03 | <.0001 |
| time | 5 | 519 | 8.25 | <.0001 |
| trtnew*time | 10 | 519 | 3.71 | <.0001 |

The Mixed Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.NEW2 |
| Dependent Variable | y |
| Covariance Structure | Unstructured |
| Subject Effect | id |
| Estimation Method | ML |
| Residual Variance Method | None |
| Fixed Effects SE Method | Model-Based |
| Degrees of Freedom Method | Between-Within |

Dimensions

| | |
|---|---|
| Covariance Parameters | 21 |
| Columns in X | 8 |
| Columns in Z | 0 |
| Subjects | 523 |
| Max Obs Per Subject | 6 |
| Observations Used | 2468 |
| Observations Not Used | 670 |
| Total Observations | 3138 |

Iteration History

| Iteration | Evaluations | -2 Log Like | Criterion |
|---|---|---|---|
| 0 | 1 | 22063.13468076 | |
| 1 | 3 | 20255.08219146 | 0.00579452 |
| 2 | 2 | 20224.45097439 | 0.00070989 |
| 3 | 1 | 20218.22109780 | 0.00004441 |
| 4 | 1 | 20217.85586897 | 0.00000028 |
| 5 | 1 | 20217.85369080 | 0.00000000 |

```
Convergence criteria met.


Estimated R Matrix for id 1

Row        Col1        Col2

  1      357.07      227.20
  2      227.20      553.82


     The Mixed Procedure

Covariance Parameter Estimates

Cov Parm    Subject    Estimate

UN(1,1)     id           357.07
UN(2,1)     id           240.70
UN(2,2)     id           442.31
UN(3,1)     id           227.20
UN(3,2)     id           410.52
```

| | | |
|---|---|---|
| UN(3,3) | id | 553.82 |
| UN(4,1) | id | 218.99 |
| UN(4,2) | id | 380.37 |
| UN(4,3) | id | 485.04 |
| UN(4,4) | id | 592.86 |
| UN(5,1) | id | 206.63 |
| UN(5,2) | id | 342.41 |
| UN(5,3) | id | 443.17 |
| UN(5,4) | id | 514.05 |
| UN(5,5) | id | 613.21 |
| UN(6,1) | id | 184.15 |
| UN(6,2) | id | 307.33 |
| UN(6,3) | id | 387.60 |
| UN(6,4) | id | 465.04 |
| UN(6,5) | id | 530.13 |
| UN(6,6) | id | 584.45 |

Fit Statistics

| | |
|---|---|
| -2 Log Likelihood | 20217.9 |
| AIC (smaller is better) | 20271.9 |
| AICC (smaller is better) | 20272.5 |

BIC (smaller is better) 20386.9

Null Model Likelihood Ratio Test

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 20 | 1845.28 | <.0001 |

Solution for Fixed Effects

| Effect | trtnew | Estimate | Standard Error | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | | 90.0941 | 0.9663 | 519 | 93.24 | <.0001 |
| trtnew | 1 | 2.3082 | 2.1578 | 519 | 1.07 | 0.2852 |
| trtnew | 2 | 1.5762 | 2.1491 | 519 | 0.73 | 0.4636 |
| trtnew | 3 | 0 | . | . | . | . |
| time | | -1.2618 | 0.1713 | 519 | -7.37 | <.0001 |
| time*trtnew | 1 | 0.4097 | 0.4057 | 519 | 1.01 | 0.3130 |
| time*trtnew | 2 | 1.7698 | 0.4404 | 519 | 4.02 | <.0001 |
| time*trtnew | 3 | 0 | . | . | . | . |

The Mixed Procedure

Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| trtnew | 2 | 519 | 0.71 | 0.4911 |
| time | 1 | 519 | 7.83 | 0.0053 |
| time*trtnew | 2 | 519 | 8.14 | 0.0003 |

## Summary

When analyzing data from the parallel groups repeated measures design, consider the following strategy.

1. Choose a working covariance structure. (Note that the choice of the mean model and the covariance structure are interdependent.) Use the REML (default) log-likelihood as the criterion for selecting a covariance structure. As a general rule of thumb, the unstructured model should be used unless a simpler covariance structure is clearly satisfactory. (**Restricted (residual) maximum likelihood (REML)**: REML is an alternative to full maximum likelihood estimation and is typically the default method in most statistical packages. Rather than maximizing the likelihood of the data, it maximizes the likelihood of the observed residuals. REML obtains initial estimates of the fixed effects using ordinary least squares and then using these estimates it maximizes the likelihood of the residuals (in which the fixed effects are subtracted off) to obtain estimates of the variance parameters. Finally the estimated variance parameters are used to obtain generalized least squares estimates of the fixed effect parameters. REML is a good alternative to

ML when the sole focus is on estimation of the variance components. The variance components obtained via ML are biased when the samples are small while REML estimates are unbiased. The problem with REML for model building is that the "likelihoods" obtained for models with different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. For this reason we have not used REML in this course. You must be aware of the distinction between ML and REML because REML is the default for most software packages. Thus you will generally need to explicitly specify maximum likelihood estimation if that's what you desire.)

2. Decide in advance whether to model the effect of treatment on patterns of change by

   (a) time by treatment interaction with time coded as a categorical variable,

   (b) time by treatment interaction with time coded as a continuous variable, or

(c) a treatment main effect ANCOVA model with baseline treated as a covariate and time treated as a categorical variable

The ML log likelihood (PROC MIXED METHOD=ML) may be used to compare any nested models differing by more than one degree of freedom.

3. Determine the final form of the regression (mean) model

4. Fit the final model using REML

5. Only estimation and test of covariate effects are considered here. It is also of great interest in testing and estimating covariance matrix.

# Lecture 20: Power & Sample Size Calculation

*Reading Assignment:*

- Muller and Fetterman, Chapter 17: "Understanding and Computing Power for the GLM" (for more background)

## Motivation

One of the most common questions asked of a statistician about study design is the number of patients to include.

It is an important question, because if a study is too small it will not be able to answer the question posed, and would be a waste of time and money. It could also be deemed unethical because patients may be put at risk with no apparent benefit.

However, it is also undesirable for studies to be too large because resources would be wasted if fewer patients would have sufficed, and it is unethical to expose more patients to the less effective treatment.

Next we will discuss the power calculation for continuous response variables followed by categorical response variables.

## Continuous Response Variable

### Model Statement

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \mathbf{e}_{n\times 1}$$

with $n >> p$. Assume $\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ and that $\boldsymbol{\beta}$ contains fixed and unknown constants. Also assume that $\mathbf{X}$ contains fixed values, known without appreciable error, conditional upon having collected the sample. Any power computed with the methods described in this chapter applies only to the particular choice of $\mathbf{X}$ used in the calculations. Changing $\mathbf{X}$ changes the design.

### Estimation

Assume $\mathbf{X}$ is of full rank. Let

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Then the predicted values can be computed as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

the sum of squares error equals

$$SSE = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'[I - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y},$$

and the variance estimate,

$$\hat{\sigma}^2 = \frac{SSE}{n-p}.$$

Let $\boldsymbol{\beta}$ be the primary parameters and $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$ be the secondary parameters. Assume $\mathbf{C}$ an $a \times p$ matrix ( $a \leq p$) with full (row) rank of $a$. Therefore

$$\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$$

is full (row) rank of $a$. Thus $\boldsymbol{\theta}$ is both estimable and testable.

## The General Linear Hypothesis

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

$$\text{versus } H_A : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

Estimability of $\boldsymbol{\theta}$ and full rank of $\mathbf{M}$ ensures uniqueness of the sum of squares for the hypothesis,

$$SSH = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

and also ensures uniqueness and existence of the likelihood ratio test.

**Test Statistic, Null Case**

Under the assumptions described above, the maximum likelihood approach provides a test with type I error rate exactly equal to the

target, $\alpha$. The likelihood ratio test statistic has many equivalent forms:

$$
\begin{aligned}
F_{obs} &= \frac{SSH/a}{SSE/(n-p)} \\
&= \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\hat{\sigma}^2} \\
&= \frac{\hat{\rho}_*^2/a}{(1-\hat{\rho}_*^2)/(n-p)},
\end{aligned}
$$

where $\hat{\rho}_*^2$ represents a generalized squared correlation. When the models spans an intercept and the hypothesis does not span the intercept, then the generalized correlation reduces to the usual multiple correlation ("corrected" for the intercept). Under $H_0$, $F \sim F(a, n-p)$.

## Test Statistic, Non-Null Case

Define *power* as the probability of rejecting $H_0$, whether or not $H_0$ is

true or not. This definition differs slightly from the traditional definition. But this new definition greatly simplifies the discussion of power. With the current definition, if $H_0$ holds, then the power of the GLH equals the type I error rate. A test is *unbiased* if the expected rejection rate is no more than $\alpha$ for null cases and no less than $\alpha$ for alternative cases. Among all tests unbiased and invariant to location and scale, the likelihood ratio test represents the uniformly most powerful test.

Define

$$
\begin{aligned}
f_A &= \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)/a}{\sigma^2} \\
&= \frac{\rho_*^2/a}{(1-\rho_*^2)/n} \\
&\approx \frac{\rho_*^2/a}{(1-\rho_*^2)/(n-p)}.
\end{aligned}
$$

The last equation has no "hats" compared to the last equation in $F_{obs}$, which uses estimates of parameters. Note that $f_A$ is a parameter, a constant, while $F_{obs}$ is a random variable.

Under $H_A$,

$$F_{obs} \sim F(a, n-p, \omega),$$

with

$$\omega = a \times f_A = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \mathbf{M}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)}{\sigma^2}$$

$$= \frac{\rho_*^2}{(1 - \rho_*^2)/n} \qquad (*)$$

Refer $\omega$ as the noncentral parameter because it captures the amount by which the model deviates from the central case.

## Properties of $\omega$ in the GLM

First note that $0 \leq \omega < \infty$. Having $\omega = 0$ implies $H_0$ holds and power equals $\alpha$. Also changing the scale of the data (such as from meters to centimeters) does not change $\omega$.

For the independent groups $T$ test of equality of means, assuming

equal cell sizes,

$$\omega = \frac{n}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma^2}.$$

For a multiple regression model that includes an intercept, $\omega$ associated with the overall test of (corrected) regression, the usual test of all slopes equal to zero is

$$\begin{aligned}\omega = \frac{\sigma_Y^2 \rho^2}{\sigma_Y^2(1-\rho^2)/n} &= \frac{\rho_*^2}{(1-\rho_*^2)/n} \\ &= \frac{n\rho_*^2}{(1-\rho_*^2)}\end{aligned}$$

Here $\sigma^2 = \sigma_Y^2(1-\rho^2)$ represents the usual residual variance for the model; while $\sigma_Y^2$ represents the variance of $\mathbf{Y}$. In turn, $\rho^2$, the usual squared multiple correlation coefficient, provides a scale free measure of effect and $(1-\rho^2)$ provides a scale free measure of residual variance.

In both special cases, $\omega$ increases with sample size, decreases with error variance, and increases with amount of effect. Equation (*) allows generalizing this statement to all cases of GLM. In the general case, $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ captures the size of the effect, while sample size $n$, hides in $\mathbf{M} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ through $\mathbf{X}$. Consider a balanced ANOVA design with cell mean coding, what is $(\mathbf{X}'\mathbf{X})^{-1}$? In summary, $\omega$ varies only as function of sample size, mean difference, and error variance.

## Computing Power

1. Specify $\alpha, \sigma^2, \mathbf{X}, \boldsymbol{\beta}, \mathbf{C}$ and $\boldsymbol{\theta}_0$.
2. Find the critical value, $f_{crit} = F_F^{-1}(1 - \alpha; a, n - p)$;.
3. Compute the noncentral parameter, $\omega$, as defined in equation (*).
4. Compute $Power = 1 - F_F(f_{crit}; a, n - p, \omega)$.

## Factors in Choosing a Design

1. Test size $\alpha$, which changes when using a Bonferroni correction for multiple analyses
2. The size of $\sigma^2$, which must often be varied due to uncertainty about the variable, the population or the study design
3. Varying $\mathbf{X}$ includes changing total sample size, cell size ratios, and the distribution of control variables. Typically balanced designs maximize power.
4. Varying $\boldsymbol{\beta}$ evaluates the impact of the strength and pattern of effects. Some choices of $\boldsymbol{\beta}$ are equivalent in terms of power: for example, for an overall test of equality of means in a three-group ANOVA balanced design, the following choices for $\boldsymbol{\beta}$ are equivalent: $[200, 210, 220]', [0, 10, 20]', [210, 200, 220]'$.
5. The choice of $\mathbf{C}$ plays the primary role in specifying the hypothesis.

6. The choice of $\boldsymbol{\theta}_0$ completes specification of the hypothesis. In most cases, $\boldsymbol{\theta}_0 = \mathbf{0}$.

**Using Parameter Estimates in Power Analysis**

Unfortunately, several quantities are required before we can do any calculations (and the argument is a little circular!) Speculation drives power analysis not data. Despite that, in some cases data from an earlier study fuel the speculation. Estimation of $\sigma^2, \boldsymbol{\beta}$ or both may be used to compute $\omega$. As a function of one or more parameter estimates, the noncentral value estimate becomes a random variable, as does the corresponding power. The process adds additional source of uncertainty.

Taylor and Muller (1995) proposed the following methods to construct confidence intervals for both noncentrality and power when using $\hat{\sigma}^2$: let

$$\hat{\omega} = \frac{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)/a}{\hat{\sigma}^2}$$

with $\hat{\sigma}^2$ based on $v$ degrees of freedom. Also let $c_{crit} = F^{-1}_{\chi^2}(\alpha_c; v)$, the $\alpha_c$ quantile of a $\chi^2$ random variable. Compute the $\alpha_c$ quantile for $\omega$ as

$$\hat{\omega}_c = \hat{\omega}\frac{c_{crit}}{v}.$$

Using $\hat{\omega}_c$ to compute power yields an $\alpha_c$ quantile for power. They reported similar results based on estimated $\sigma^2$ and $\boldsymbol{\beta}$.

**Example1: Kidney Disease**

Falk et al (1992) rabdomly assigned 24 participants to one of two treatments intended to slow the worsening of kidney disease. Higher levels of creatinine indicate worse function. Using the reciprocal of serum creatinine level as the dependent variable allowed the investigators to meet the Gaussian assumption. The scientists considered an increase of 0.50dL/mg a clinically important improvement.

If we state the model as

$$\begin{aligned}\mathbf{y}_{24} &= \mathbf{X}_{24\times 2}\boldsymbol{\beta}_{2\times 1} + \mathbf{e} \\ &= \begin{bmatrix} \mathbf{1_{12}} & 0_{12} \\ 0_{12} & \mathbf{1_{12}} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \mathbf{e}\end{aligned}$$

The hypothesis of interest uses $\mathbf{C} = [-1, 1]$ and $\boldsymbol{\theta}_0 = 0$. Let $\delta = \mu_2 - \mu_1$. Then $\boldsymbol{\theta} = \delta = 0.5$, which reflects the investigators' interest. Assuming $\sigma^2 = 0.068$ that estimated from the study, which led to a $\hat{\omega} = 22.06$ and a power of 0.96 under $\alpha = 0.01$. Taylor and Muller computed the 95% CI for $\omega$ as [11.01, 36.880] and for power as [0.688, 0.999]. The asymmetry of the $\chi^2$ leads to asymmetric confidence intervals. Taylor and Muller recommended using one-sided confidence intervals for power, since we usually wish to make statements of inequality, such as ensuring power no less than $P$. The one sided interval is [0.75, 1].

**Example2: Medical Cost** Suppose we want to design a study to

determine whether there is a linear relationship between nursing home patients' ages and their annual costs for medication. It would be considered unimportant from an economic and medical standpoint if age explained less than .04 of the variability in medication cost. We want a significance level of $\alpha = 0.05$. Calculate the power when sample size n=200.

Solution: $f_{crit} = F_F^{-1}(1-\alpha; a, n-p) = F_F^{-1}(0.95; 1, 198) = 3.889$

$\omega = \frac{n\rho^2}{(1-\rho^2)} = 200 * 0.04/(1-0.04) = 8.33$

$power = 1 - F_F(f_{crit}; 1, n-p, \omega) = 1 - F_F(3.889; 1, 198, 8.33) = 0.82.$

## Power Reporting

Single power values rarely suffice to inform the scientist. As statisticians, the authors use tables and plots of power values to return the choices of sample size to the principal investigator. A typical table involves varying two or three of the factors controlling power: mean difference, variance and sample size.

Power for comparing two means

| Sample Size per group | D etectable Effect Size | |
|---|---|---|
| | 80% power | 90% power |
| 10 | 1.3 | 1.5 |
| 20 | 0.9 | 1.1 |
| 50 | 0.6 | 0.7 |
| 100 | 0.4 | 0.5 |
| 200 | 0.28 | 0.33 |

where effect size$= \frac{|\mu_1 - \mu_2|}{\sigma}$.

## Dichotomous Responses

For a binary outcome we need to specify type I error, and proportions P1 and P2 where P1 is the expected outcome under the control intervention and P1- P2 is the minimum clinical difference which it is worthwhile detecting.

1. Specify $\alpha, P1$ and $P2$.
2. Find the critical value, $z_{crit} = N^{-1}(1 - \alpha/2)$;
3. Calculate $\omega = \frac{P1-P2}{\sqrt{P1(1-P1)/n1+P2(1-P2)/n2}}$
4. Compute $Power = 1 - N(z_{crit}; \omega, 1)$.

Power for comparing two proportions

| P1 | P2 | n1 | n2 | Power |
|---|---|---|---|---|
| 0.2 | 0.3 | 100 | 100 | 0.38 |
| 0.2 | 0.3 | 200 | 200 | 0.64 |
| 0.15 | 0.3 | 100 | 100 | 0.73 |
| 0.15 | 0.3 | 200 | 200 | 0.95 |

## Illustrate Power by Figure

Power of differential expression study for three different p-value cutoffs.

**Software:**

- Commercial one: nquery
  https://uncapps.its.unc.edu/vpn/index.html
- SAS and R have various functions for power analysis
- online power calculation: http://powerandsamplesize.com/

## How Much to Do?

In practice the sample size is often fixed by other criteria, such as finance or resources, and the formula is used to determine a realistic effect size. If this is too large, then the study will have to be abandoned or increased in size.

Five key questions regarding sample size:

1. What is the main purpose of the study?

2. What is the principal measure of patient outcome?

3. How will the data be analyzed to detect a treatment difference?

4. What type of results does one anticipate with standard treatment?

5. How small a treatment difference is it important to detect and with what degree of certainty?

Thus in order to calculate the sample size for a study it is first necessary to decide upon what your outcome is. If your outcome variable is continuous you will need to have some measure of what you would expect its mean value to be in the control group together with an estimate of its standard deviation. You will also need to know what size of effect you expect or is desirable (be realistic with this). If your outcome variable is binary you will need to have an idea of the proportions falling into the two outcome categories, and what change in these proportions can be expected or is desirable.

After deciding on the purpose of the study and the principle outcome measure, the investigator must decide how the data are to be summarized and analyzed to detect a treatment difference. Thus, the

investigator must choose an appropriate summary measure of this outcome and then calculate a sample size based on the smallest treatment difference in this summary measure that is of such clinical value that it would be very undesirable to fail to detect. Given answers to all of the five questions above, we can then calculate a sample size.

# Summary

Topics not in this summary will not be tested on.

Finally, the end of the semester! We made it!

## Topic 1: Introduction and Overview

1. Why linear regression, why not t-test?
2. Basic concepts: Population, Sample, parameter, statistic
3. Statistical Activities: Parameter Estimation, Inference

## Topic 2: Linear Algebra Review

1. Matrix operation, matrix addition, matrix multiplication ...
2. An *orthogonal matrix* is a **square matrix** with $\mathbf{A}' = \mathbf{A}^{-1}$.
3. Rules of Matrix Operation.
4. Linear Dependence and Rank, matrix determinant [this is important but will not come up on exam]
5. Positive Definite and Semi-positive Definite Matrices [important but will not come up on exam]
6. Inverse and Generalized Inverse [she will ask for inverse not generalized inverse]
7. Eigenvalues, Eigenvectors. Suppose $\mathbf{A}$ is an symmetric matrix. Then there exists an orthogonal (column orthonormal) matrix $\mathbf{V}$ such that $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$.

8. <u>Random Vectors and Matrices</u>

$$E(\mathbf{AY} + \mathbf{b}) = \mathbf{A}E(\mathbf{Y}) + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$
$$\text{Cov}(\mathbf{AY} + \mathbf{b}) = \mathbf{A}\text{Cov}(\mathbf{Y})\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'.$$

9. <u>Important Distributions for Linear Models</u>. If $Z \sim N(0, 1)$, $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$, and $X_1$ and $X_2$ are independent. Construct random variables following t-distribution and F distribution. If given these three variables, should know how to construct F and t-stats.
t = Z/sqrt(X1 OR X2).

10. Maximum Likelihood Estimates (MLE)

## Topics 3 and 4: Simple Linear Regression and the General Linear Model: Estimation and Testing

1. $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\ \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$
2. Least Squares Estimation: $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
   $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and Cov$(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
3. HILE Gauss
   - Existence Assumption can't check; assume
   - Linearity Assumption can check to some degree
   - Independence Assumption can't check; assume
   - Homogeneity Assumption should know meaning and how to check with real data
   - Gaussian Errors Assumption important for small sample size. lenient for larger data.
4. $\boldsymbol{\beta}$ is the vector of primary parameters, and $\boldsymbol{\theta}_{a\times 1} = \mathbf{C}_{a\times p}\ \boldsymbol{\beta}_{p\times 1}$ is

a vector of secondary parameters, defined by $\mathbf{C}$, the *contrast matrix*. Each row of $\mathbf{C}$ defines a new scalar parameter in terms of the $\boldsymbol{\beta}$'s, e.g.,$\beta_1 - \beta_2$. The general linear hypothesis is

$$\begin{aligned} H_0 : \boldsymbol{\theta}_{a\times 1} &= \boldsymbol{\theta}_0 \\ H_A : \boldsymbol{\theta}_{a\times 1} &\neq \boldsymbol{\theta}_0. \end{aligned}$$

5. Estimability and Testability of a Parameter. If $\mathbf{X}$ is full rank , then $\widehat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable, and any (nonzero) $\boldsymbol{C}$ gives estimable $\boldsymbol{\theta}$. If $\boldsymbol{C}$ is full rank, $\boldsymbol{\beta}$ is testable.

6. Computation of Test Statistic and p-value. Let $\mathbf{M}_{a\times a} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ and $SSH_{1\times 1} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. The test-statistic is

$$F_{obs} = \frac{SSH/a}{SSE/(n-p)} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\widehat{\sigma}^2} = \frac{MSH}{MSE}$$

bottom part is MSE (aka sigma_hat^2)

## Topic 5: Some Distributional Results for the GLM

- If $\mathbf{X}$ is full rank, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $\boldsymbol{\theta} = \mathbf{C}_{a\times p}\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\theta}} \sim N_a(\boldsymbol{\theta}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')$.
- Predicted Values: Conditional Means and Future Observations
  - $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{H}\mathbf{y}$,
  - $E(\widehat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$,
  - $\text{cov}(\widehat{\mathbf{y}}) = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Definitions and Properties of Residuals
- Residual Variance $\widehat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-p} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}$

## Topic 6: Multiple Regression: General Consideration

- Basic Sum Squares:

$$USS(\text{total}) = USS(\text{model}) + SSE, \quad \mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}.$$

$$\begin{aligned} CSS(\text{total}) &= CSS(\text{model}) + SSE \\ \mathbf{y}'\left[\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} &= \mathbf{y}'\left[\mathbf{H} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

- $$F_{obs} = \frac{MS(\text{hypothesis})}{MSE} = \frac{SSH/dfH}{SSE/dfE}$$

This is the other formula. this will be useful given two ANOVA tables (one based on full and one based on reduced. Won't need M and C any more.

$$= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})}$$

$$= \frac{CSS(\text{Regression})/(p-1)}{SSE(full)/(n-p)}.$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1-\alpha,\, p-1,\, n-p) = f_{crit}$.

The usual test of overall regression assumes model spans an intercept and excludes the intercept from the test.

- ANOVA table.
- Usual "Corrected" $\boldsymbol{R^2}$:
  $R_{\mathsf{c}}^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression})+SSE(\text{full})} = \frac{CSS(\text{Regression})}{CSS(\text{total})}$. $R_c^2$ estimates $\rho_{\mathsf{c}}^2$, the population ratio of model to total variance, with $0 \leq \rho_{\mathsf{c}}^2 \leq 1$ and $0 \leq R_{\mathsf{c}}^2 \leq 1$.
- The corrected test for overall regression,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

holds if and only if $H_0 : \rho_{\mathsf{c}}^2 = 0$

This was an important fact for one question on the midterm.

## Topic 7: Testing Hypotheses in Multiple Regression

- All tests compare two models: the full model and the reduced model (this is the basic idea of likelihood ratio tests, called the *likelihood ratio principle*).

- Overalltest: $F_{obs} = \dfrac{CSS(\beta_1,\ldots,\beta_{p-1})\big/(p-1)}{SSE(\beta_0,\ldots,\beta_{p-1})\big/(n-p)}$.

- Added-Last Test: the *added-last test* seeks to assess the usefulness of one predictor, above and beyond all others. Coefficient Estimates/t-test table, Type III table. The F statistic is

Need to know Type I vs. Type III. Type III is added-last test. df always 1 for numerator b/c only testing 1 parameter. Look at the corresponding C-matrix below.

$$F_{obs} = \frac{\dfrac{SSE(\text{reduced}) - SSE(\text{full})}{dfE(\text{reduced}) - dfE(\text{full})}}{SSE(\text{full})/dfE(\text{full})} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)'\mathbf{M}^{-1}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/dfH}{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}/dfE},$$

where $\mathbf{C} = \begin{bmatrix} 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \end{bmatrix}_{1\times p}$

- Added-in-Order Test: the *added-in-order test* seeks to assess the

Type I test. P-value changes based on order of putting things in model, unlike Type III.

contribution of predictor $j$ above and beyond all of the preceding $j-1$ predictors (without the $j+1$, $j+2$, etc.predictors in the model).

- Group Added-Last Tests
- Group Added-in-order Tests

## Topic 8: Correlations

population correlation = $$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

sample correlation = $$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)}}.$$

Partial correlations describe the strength of the linear relationship between two variables, $Y$ and $X$, after controlling for the effects of other variables $\mathbf{Z}$.

We won't be tested over partial correlations.

## Topic 9: GLM Assumption Diagnostics

- The First Step: Get to Know Your Data
- Homogeneity: violations seen in the pattern of residuals.
- Independence: assessed through logic of sampling scheme.
- Linearity: examine pattern of residuals.
- Existence: (finite sample...).
- Gaussian distribution: distributional assessment involves box plot of residuals, histogram of residuals, and test of Gaussian distribution of residuals. (The discrepancy between $T$ and Gaussian random variables somewhat inflates the probability of rejecting the null...why?)
- Outliers: leverage, Influence: Cook's Distance

## Topic 10: Computation Diagnostics

- Colinearity
- Eigenanalysis
- Condition Number and Condition Index: the *condition index* for the $k$th eigenvalue equals $\sqrt{\lambda_1/\lambda_k}$. The maximum condition index, called the *condition number*
- $\boldsymbol{R_j^2}$, Tolerance, and VIF
  $R_j^2 = R^2(X_j, \{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\})$

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}}.$$

- Leverage
- Cook's distance

## Topic 11: Selecting the Best Model

1. Specify the maximum model under consideration.
2. Specify a criterion for model selection.
3. Specify a strategy for applying the criterion.
4. Conduct the analysis.

# Topic 12: ANOVA

- Coding schemes

$$\text{Es}(\mathbf{X}_{ref}) = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3\times 3} \qquad \text{Es}(\mathbf{X}_{cell}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}$$

$$\text{Es}(\mathbf{X}_{anova}) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}_{3\times 4}$$

$$\text{Es}(\mathbf{X}_{effect}) = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}_{3\times 3}$$

- Step down test

## Topic 13: Coding Schemes for Regression

- (Regression) $\quad \boldsymbol{y} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mathbf{1} & \mathbf{x} \\ \mathbf{1} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad + \boldsymbol{\varepsilon}$

- (ANOVA) $\quad \boldsymbol{y} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad + \boldsymbol{\varepsilon}$

- (Intercept Only) $\quad \boldsymbol{y} = \begin{bmatrix} \mathbf{1} \\ \mathbf{1} \\ \mathbf{1} \end{bmatrix} [\beta_0] \quad + \boldsymbol{\varepsilon}$

- (Null) $\quad \boldsymbol{y} = \boldsymbol{\varepsilon}$

# Topic 14: Logistic Regression

We will not be tested on Poisson!
But we will be on Logistic.

- Definition of odds, and odds ratio
- The general logistic regression model is given by

$$\begin{aligned} \text{logit}(p_i) &= \log\left(\frac{p_i}{1-p_i}\right) \\ &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} \end{aligned}$$

  with $y_i \sim \text{Bernoulli}(p_i)$, $i = 1, \ldots, n$, and the $y$'s independent of each other.
- Interpretation of regression coefficients in terms of odds ratio.
- Model comparison by likelihood ratio test Don't worry about
- Logistic regression with categorical covariates and their interactions.
- Goodness of fit test Don't worry about

# Topic 15: Mixed Effects Model

Everything about regression models same for mixed effects model. Need to know how to interpret beta from output. Only thing different here is covariates. (e.g. "Given ANOVA table, which covariate structure fits the data better?")

- When data are correlated and the independence assumption does not hold, mixed effects models are one way to adjust for the non-independence of observations
- Random effects may be introduced to account for the fact that observations within one subject (or more generally, within one cluster) may be more alike than observations from different clusters
- Forms of covariance matrices for clustered and repeated measurements
- Parameter interpretation of models for longitudinal data

$$USS(\text{total}) = \boldsymbol{y}'\boldsymbol{y} = \sum_{i=1}^{n} y_i^2.$$

$$USS(\text{model}) = \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} = \sum_{i=1}^{n} \widehat{y}_i^2.$$

$$USS(\text{total}) \quad = \quad USS(\text{model}) + SSE$$

$$CSS(\text{total}) \quad = \quad USS(\text{total}) - SSI$$

$$= \quad \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

$$CSS(\text{model}) \quad = \quad USS(\text{model}) - SSI$$

$$= \quad \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2.$$

$$CSS(\text{total}) \quad = \quad CSS(\text{model}) + SSE$$