Statistics One

Introduction

Introduction

- Statistics, broadly defined, is a scientific discipline devoted to the study of data

Introduction

- *Data*: a collection of numbers assigned as values to quantitative variables and/or characters assigned as values to qualitative variables

Introduction

- Example: Academic records of children in elementary school

Introduction

Student	Gender	Age (Months)	Math	History	Language
TR	M	80	95	91	73
CS	F	79	79	75	87
PP	M	84	75	82	84
DB	M	84	94	98	95
MM	F	82	93	78	78
AC	M	83	91	79	80
…					

Introduction

- *Data*: the lowest level of abstraction from which information and then knowledge are derived
    - Data → Information → Knowledge

Introduction

- *Statistician*: a person who is skilled in applying the tools of Statistics

Introduction

- Types of Statisticians
  - Academic research
  - Medical research
  - Survey studies
  - Education
  - Market research
    - Analytics and Big Data

Introduction

- *Statistic*: a quantity calculated from a sample of data
    - Average Age of students
    - Average Math grade
    - Standard deviation of Math grade

Introduction

- *Sample*: a subset of the population
- *Population*: the entire collection of cases to which we want to generalize

Introduction

- *Statistic*: a numerical measure that describes a characteristic of a sample

- *Parameter*: a numerical measure that describes a characteristic of a population

Introduction

- *Descriptive statistics*: procedures used to summarize, organize, and simplify data

- *Inferential statistics*: procedures that allow for generalizations about population parameters based on sample statistics

Introduction

- Research methods
  - Descriptive
  - Correlational
  - Experimental

Introduction

- Descriptive
  - Organize and summarize the data

Introduction

- Correlational
  - Examine relationships among variables
    - Is Math grade correlated with History grade?

Introduction

- Experimental
  - Randomly assign students to different schedules
    - Year-round
    - Summer break
    - Is achievement affected by schedule?

Introduction

- The International Year of Statistics, 2013!
  - For information:
    - www.statistics2013.org

Introduction

- “Statistics is becoming more critical as academia, businesses, and governments come to rely on data-driven decisions, greatly expanding the demand for statisticians.”

END INTRODUCTION

Statistics One

Lecture 1
Experimental Research

1

Three Segments

- Example 1: Polio Vaccine
- Example 2: Memory Training
- The concept of random

2

Lecture 1 ~ Segment 1

Example 1: Polio Vaccine

3

Polio Vaccine

- In the first half of the 20th century there were approximately 20,000 cases of polio per year in the USA
- In 1952, there were 58,000 cases

4

Polio Vaccine

- In 1952, the first effective polio vaccine was developed by Dr. Jonas Salk
  - How do we know that it was effective?
    - Experimental research!
    - Randomized Controlled Experiments

5

Polio Vaccine

- Sample
  - Initial
    - 4,000 children from Virginia
  - Final
    - 1.8 million children from 44 states
- Population
  - All children in the USA

6

Polio Vaccine

- Independent variable
  - Treatment
    - Vaccine
    - Placebo
- Dependent variable
  - Polio diagnosis (measure of an individual child)
  - Rate of polio (measure of a group of children)

7

Polio Vaccine

- Double-blind experiment
  - Experimenter did not know if the treatment was vaccine or placebo
  - Child (and parents) did not know if the treatment was vaccine or placebo

8

Polio Vaccine

- Results
  - Rate (per 100,000)
    - Treatment: 28
    - Control: 71

9

Polio Vaccine

- By 1994 polio had been completely eradicated from all the Americas

10

Segment Summary

- The major benefit of randomized experiments is they allow for strong claims about causality
  - Why stuff happens!
    - Predict stuff
    - Prevent bad stuff
    - Promote good stuff

11

Segment Summary

- Strong causal claims require:
  - True independent variables
  - Random and representative samples
  - No confounds (impossible, but we try our best)

12

END SEGMENT

13

Lecture 1 ~ Segment 2

Example 2: Memory Training

14

Memory Training

- Is it possible for adults to enhance their intelligence by training their working memory?
  - Promote good stuff!

15

Memory Training

- Sample
  - College students
- Population
  - Healthy adults

16

Memory Training

- Independent variable
  - Training
    - Memory training
    - No training
- Dependent variable
  - Gain in score on an intelligence test
    - IQ gain

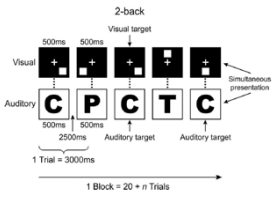
17

Memory Training

- Procedure
  - Treatment group engaged in memory training for a half hour every day for weeks
    - See next slide
  - IQ
    - All subjects completed a test of intelligence before and after training

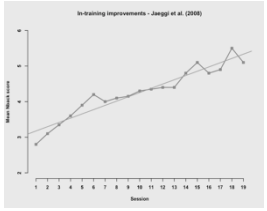
18

Memory Training



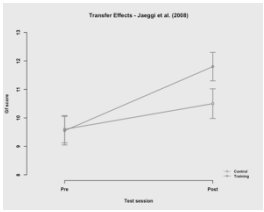
19

Memory Training



20

Memory Training



21

Memory Training

- Does it really work?
  - Potential confounds?

22

Segment Summary

- The major benefit of randomized experiments is they allow for strong claims about causality
  - Why stuff happens!
    - Predict stuff
    - Prevent bad stuff
    - Promote good stuff

23

Segment Summary

- Strong causal claims require:
  - True independent variables
  - Random and representative samples
  - No confounds (impossible, but we try our best)

24

END SEGMENT

25

Lecture 1 ~ Segment 3

The concept of random

26

Random

- Experimental research requires:
  - Random selection
  - Random assignment

27

Random

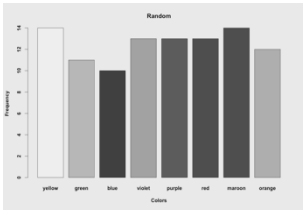
- Random selection
  - Individuals included in a sample should be randomly selected from the population

28

Illustration

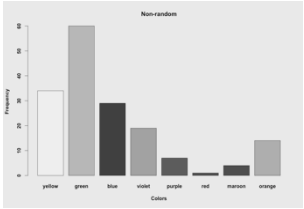
29

Random



30

Not random



31

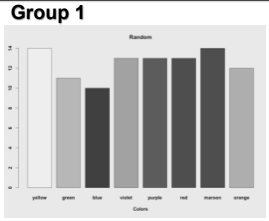
Random

- Random assignment
  - Individuals are randomly assigned to conditions

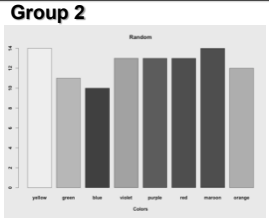
32

Random assignment

Group 1



Group 2



33

Segment Summary

- Experimental research requires:
  - Random selection
  - Random assignment

34

END SEGMENT

35

END LECTURE 1

36

Statistics One

Lecture 2
Correlational Research

1

Three Segments

- Example 1: Personality
- Example 2: Intelligence
- Example 3: Sports-related concussion

2

Lecture 2 ~ Segment 1

Example 1: Personality

3

Personality

- "As any parent of more than one child knows, children are not indistinguishable lumps of raw material waiting to be shaped. They are little people, born with personalities." – Steven Pinker

4

Personality

5

Personality

- Personality traits
  - Traits are considered to be relatively stable, distinguishable qualities of a person

6

Personality

- The Big Five personality traits
  - Openness
  - Conscientiousness
  - Extraversion
  - Agreeableness
  - Neuroticism
  - OCEAN

7

Personality

- Survey questions
  - Examples of questions to measure extraversion
    - Q1: I am the life of the party
    - Q2: I don't mind being the center of attention

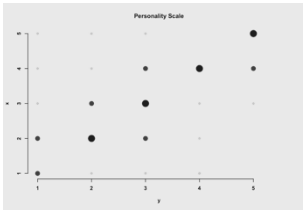
8

Personality

	Disagree				Agree
I am the life of the party.	○	○	○	○	○
I feel little concern for others.	○	○	○	○	○
I am always prepared.	○	○	○	○	○
I tend to manipulate others to get my way.	○	○	○	○	○
I get stressed out easily.	○	○	○	○	○
I have a rich vocabulary.	○	○	○	○	○
I tend to lack remorse.	○	○	○	○	○
I don't talk a lot.	○	○	○	○	○
I am interested in people.	○	○	○	○	○
I leave my belongings around.	○	○	○	○	○
I tend to want others to admire me.	○	○	○	○	○
I am relaxed most of the time.	○	○	○	○	○
I have difficulty understanding abstract ideas.	○	○	○	○	○

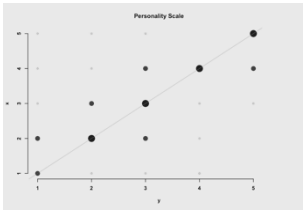
9

Personality



10

Personality



11

Personality

- The theory of five components of personality is supported by correlational research
  - Surveys, interviews, & observations of behavior
  - For more information, see
    - Five Factor Model of Personality

12

END SEGMENT

13

Lecture 2 ~ Segment 2

Intelligence

14

Intelligence

- *Intelligence*: "A very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings—"catching on," "making sense" of things, or "figuring out" what to." – Wall Street Journal, 1994

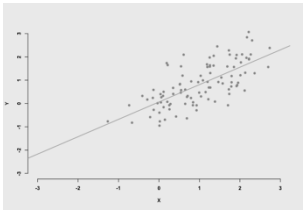
15

Intelligence

- Theories of intelligence have been proposed based on detailed analysis of patterns of correlations across different types of tests
  - I refer to these as "studies" of intelligence rather than "experiments" because no variable is manipulated (no independent variable)

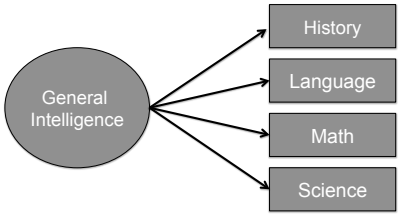
16

Positive correlation



17

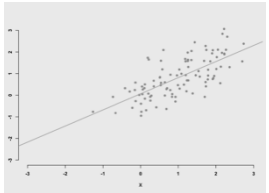
Theory of intelligence



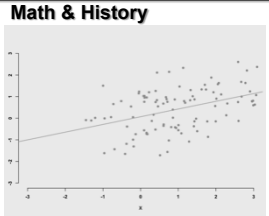
18

Pattern of correlations

Math & Science

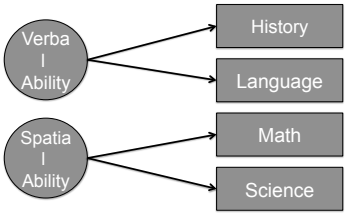


Math & History



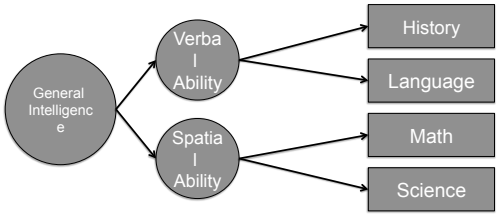
19

Theory of intelligence



20

Theory of intelligence



21

Intelligence

- The hierarchical model of intelligence is supported by correlational research
  - There is a general ability, and
  - Several more specific abilities
  - For more information, see
    - Cattell-Horn-Carroll theory of intelligence

22

END SEGMENT

23

Lecture 2 ~ Segment 3

Sports-related concussion

24

Effects of concussion

- Sports-related concussions, especially in American football, are common and may cause neural damage and cognitive deficits

25

Effects of concussion

- Quasi-independent variable
  - Treatment
    - Suffered a sports-related concussion
    - Control group
- Dependent variable
  - Neural measures
  - Cognitive measures

26

Effects of concussion

Concentration

"I am going to read you a string of numbers and when Im done you repeat them back to be in reverse order. For example if I say 7-1-9 you would say 9-1-7"

If correct go to the next string length. If incorrect, read trial 2. 1 point for each string length. Stop after incorrect on both trials. Digits should be read at rate of 1 digit /sec

Digits Backward:

			Alternative digit lists		
4-9-3	0	1	6-2-9	5-2-6	4-1-5
3-8-1-4	0	1	3-2-7-9	1-7-9-5	4-9-6-8
6-2-9-7-1	0	1	1-5-2-8-6	3-8-5-2-7	6-1-8-4-3
7-1-8-4-6-2	0	1	5-3-9-1-4-8	8-3-1-9-6-4	7-2-4-8-5-6

27

Effects of concussion

- Confounds?
  - Prior concussions
  - Prior hits to the head (not necessarily concussions)
  - Personality types more likely to be aggressive

28

Effects of concussion

- Quasi-independent variable
  - Since the IV does not involve random and representative sampling, arguments about causality are not as strong

29

Lecture Summary

- Important concepts
  - Correlational research / Experimental research
  - I will refer to examples of correlational research as "studies" and examples of experimental research as "experiments"

30

Lecture Summary

- Important concepts
  - Many theories, especially of personality and intelligence, have been tested by investigating patterns of correlations obtained from observational studies
  - Some things simply can't be experimentally manipulated, for example, concussions! Hence, quasi-independent variables

31

END SEGMENT

32

END LECTURE 2

33

Statistics One

Lecture 3
Variables, Distributions, & Scales

1

Three segments

- Variables
- Distributions
- Scales

2

Lecture 3 ~ Segment 1

Types of variables

3

Variables

- Variables can take on multiple values
- In contrast, a constant has only one value

4

Apples and gravity

5

Variables

- The size, shape, weight, and type of apple are all variables
- Gravity, or gravitational force, is a constant on Earth

6

Types of variables

- Nominal
- Ordinal
- Interval
- Ratio

7

Stevens (1946)

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

8

Types of variables

- Nominal variables
  - Used to assign individual cases to categories
    - For example, Coursera students come from many different countries
    - *Country of Origin* is a nominal variable

9

Types of variables

- Ordinal variables
  - Used to rank order cases
    - For example, countries may be ranked according to overall population
    - *Ranking* is an ordinal variable

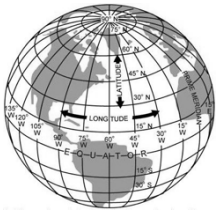
10

Types of variables

- Interval variables
  - Used to rank order cases and the distance, or interval, between each value is equal
    - For example, each country has a longitude and latitude
    - *Longitude* and *Latitude* are interval variables

11

Types of variables



12

Types of variables

- Ratio variables
  - The same as interval variables but they have a "true zero"
    - For example, *Population* (Population = 0 = extinct)
    - For example, *Age* (Age = 0 literally means NO age)
    - For example, *Temperature K°* (the Kelvin scale)

13

Stevens (1946)

TABLE 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

14

Types of variables

- Preview of variables in this course
  - Nominal variables
    - Independent variables in experimental research
      - For example, treatment to prevent polio (vaccine, placebo)
    - Quasi-independent variables in correlational research
      - For example, gender (female, male)

15

Types of variables

- Preview of variables in this course
  - Interval and Ratio variables
    - Dependent variables in experimental research
      - For example, rate of polio in a community
    - Measured variables in correlational research
      - For example, intelligence test scores

16

Types of variables

- Preview of variables in this course
  - Discrete vs. continuous variables
    - Nominal variables are discrete (categorical)
    - Interval and ratio variables are continuous
    - Ordinal variables are technically discrete but they are often treated as continuous in statistical analyses (more on this later)

17

Segment summary

- Types of variables
  - Nominal
  - Ordinal
  - Interval
  - Ratio

18

END SEGMENT

19

Lecture 3 ~ Segment 2

Distributions: Histograms

20

Histograms

- A histogram is a type of graph used to display a distribution

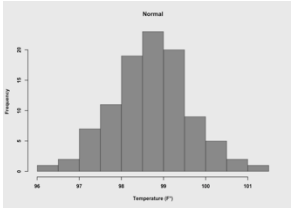
21

Histograms

- Why start with histograms?
  - To overcome the natural tendency to rely upon summary information, such as an average

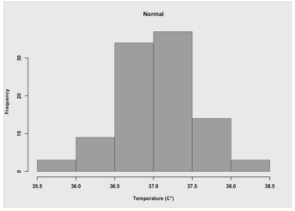
22

An example: Body temperature



23

An example: Body temperature



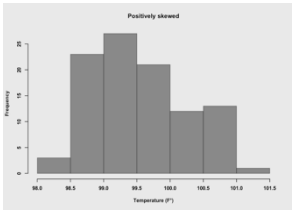
24

Histograms

- Histograms can reveal information not captured by summary statistics
  - Suppose a few children in a school are sick with influenza (flu) and have a high temperature
    - The distribution will be positively skewed

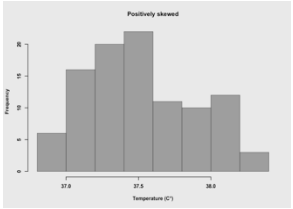
25

An example: Body temperature



26

An example: Body temperature



27

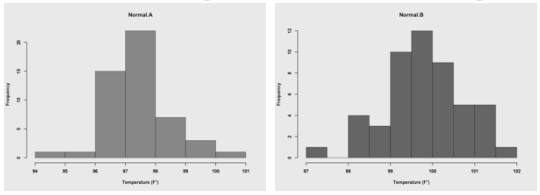
Histograms

- Not all distributions are normal
  - Suppose one group of children had the flu a week prior to a second sick group of children
  - Assume the first group received antiobiotics, which temporarily caused their body temperatures to be slightly below normal, while the second group was still above normal

28

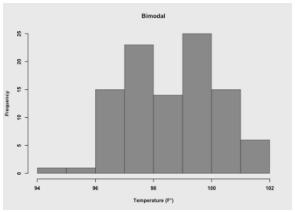
An example: Body temperature

Normal, below average **Normal, above average**



29

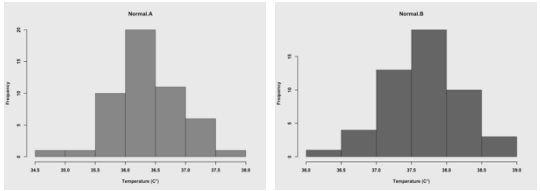
An example: Body temperature



30

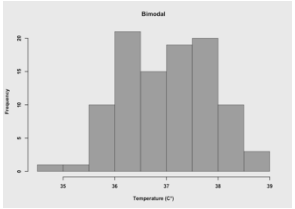
An example: Body temperature

Normal, below average **Normal, above average**



31

An example: Body temperature



32

Histograms

- Not all distributions are normal
  - Simply viewing a histogram often reveals whether a distribution is normal or not normal
  - However, sometimes it is hard to determine
    - Summary statistics help in such cases

33

Histograms

- Not all distributions are normal
  - As you view more and more distributions you will get a better sense of what is normal and what is not normal
  - So, let's look at more distributions

34

Wine tasting!

35

An example: Wine ratings

- Suppose that 100 wine experts rated the overall quality of 8 different wines on a scale of 1 to 100
  - Higher scores indicate higher quality

36

An example: Wine ratings

- Suppose four countries submitted two wines each, one red and one white
  - Argentina
  - Australia
  - France
  - USA

37

An example: Wine ratings

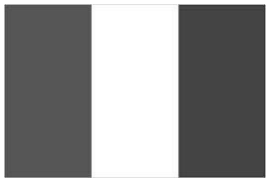
Malbec & Chardonnay

Shiraz & Pinot Grigio

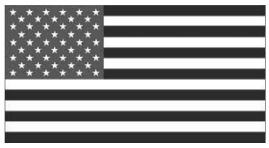
38

An example: Wine ratings

Bourdeaux & Sauvignon Blanc



Cabarnet & Reisling



39

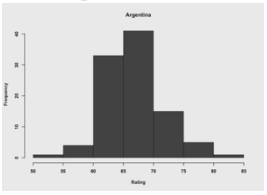
An example: Wine ratings

- Preview
  - The ratings of the red wines are normal
  - The ratings of the whites are not normal

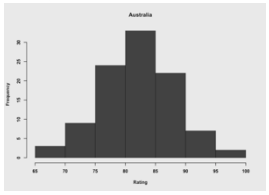
40

An example: Wine ratings

Red, Argentina



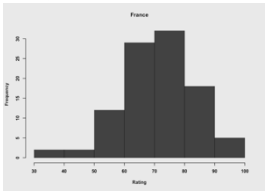
Red, Australia



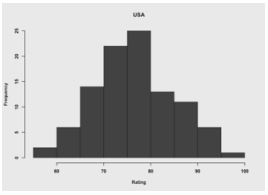
41

Four histograms

Red, France



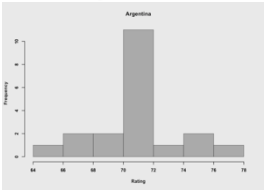
Red, USA



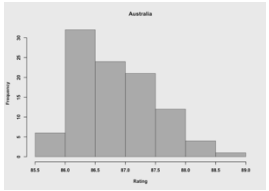
42

An example: Wine ratings

White, Argentina



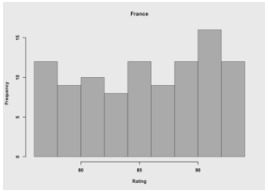
White, Australia



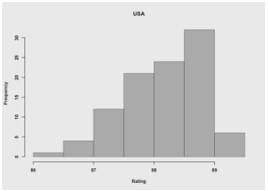
43

An example: Wine ratings

White, France



White, USA



44

Segment summary

- Histograms are used to display distributions
- Many distributions are normal

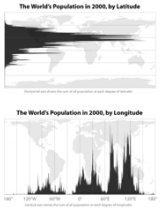
45

Segment summary

- Some distributions are not normal, for example:
  - Bi-modal
  - Positively skewed
  - Negatively skewed
  - Uniform (platykurtic)
  - Leptokurtic

46

Advanced graphs

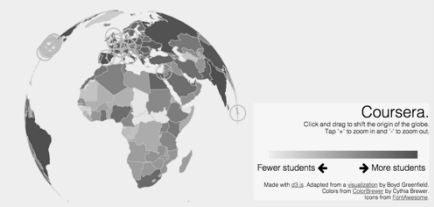


47

Advanced graphs

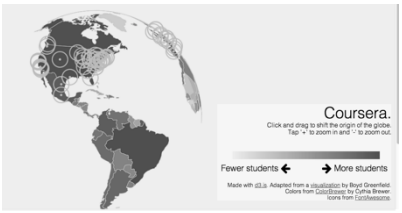
48

Advanced graphs



49

Advanced graphs



50

END SEGMENT

51

Lecture 3 ~ Segment 3

Scales of measurement

52

Scales

- Scales of measurement
  - For example, in the last segment body temperature was presented in both Fahrenheit and in Celsius
    - Different scales but both measure temperature
    - F° can be converted to C° and vice-versa

53

Scales

- In statistics, there is a standard scale
  - The Z scale
- Any score from any scale can be converted
  - To Z scores
- Allows for efficient communication

54

Z scores

- Z = (X – M) / SD
  - X is a score on an original scale (raw score)
  - M is the mean
  - SD is the standard deviation

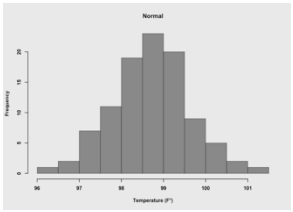
55

Z scores

- Z = (X – M) / SD
  - The mean Z-score is Z = 0
  - Positive Z scores are above average
  - Negative Z scores are below average

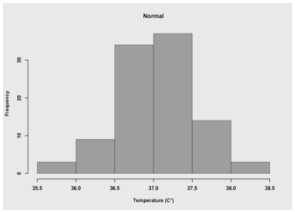
56

Body temperature F°



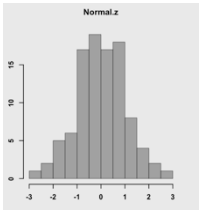
57

Body temperature C°



58

Body temperature Z



59

Z scores

- For example, assume M = 98.6, SD = .5
- Suppose an individual, X = 99.6
- Convert X to Z

60

Z scores

- Convert X to Z
  - Z = (X – M) / SD
  - Z = (99.6 – 98.6) / .5 = 2
  - Z = 2

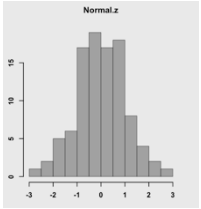
61

Percentile rank

- Percentile rank
  - The percentage of scores that fall at or below a score in a distribution
    - Assume a normal distribution
    - If Z = 0 then the percentile rank = 50th
    - 50 percent of the distribution falls below the mean

62

Body temperature Z



63

Segment summary

- The Z-scale is the standard scale in statistics
- Raw scores can be converted to Z-scores
- Z-scores can be used to find percentile rank
  - Raw score ~ Z-score ~ Percentile rank

64

END SEGMENT

65

END LECTURE 3

66

Statistics One

Lecture 4
Summary Statistics

1

Two segments

- Measures of central tendency
- Measures of variability

2

Lecture 4 ~ Segment 1

Measures of central tendency

3

Wine tasting!

Example: Wine ratings

- Suppose that 100 wine experts rated the overall quality of different wines on a scale of 1 to 100
  - Higher scores indicate higher quality

Example: Wine ratings

- Consider the red wines, which country had the highest average (mean) rating?

Example: Wine ratings (Reds)

Country	Mean = M = (ΣX) / N
Argentina	66.73
Australia	81.76
France	70.97
USA	76.38

Example: Wine ratings

- Now consider the white wines, which country had the highest average (mean) rating?

Example: Wine ratings (Whites)

Country	Mean = M = (ΣX) / N
Argentina	71.20
Australia	86.81
France	85.90
USA	88.62

Example: Wine ratings

- The mean is a measure of central tendency

Measures of central tendency

- *Measure of central tendency*: A measure that describes the middle or center point of a distribution
  - A good measure of central tendency is representative of the distribution

Measures of central tendency

- *Mean*: the average, M = (ΣX) / N
- *Median*: the middle score (the score below which 50% of the distribution falls)
- *Mode*: the score that occurs most often

Measures of central tendency

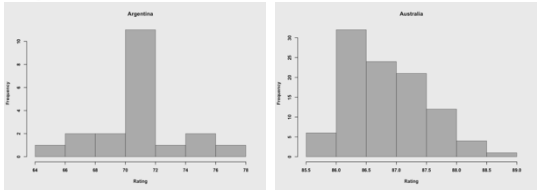
- Mean (average) is the best measure of central tendency when the distribution is normal
  - Red wine ratings
  - Another example: Grade Point Average (GPA)

Measures of central tendency

- Median (middle score) is preferred when there are extreme scores in the distribution
  - White wine ratings?
  - Another example: Household income in USA

Measures of central tendency

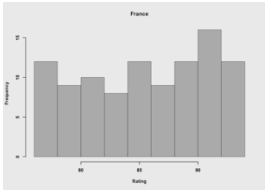
Argentina **Australia**

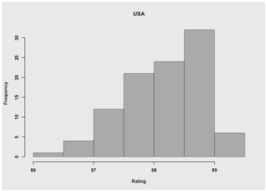


15

Measures of central tendency

France **USA**



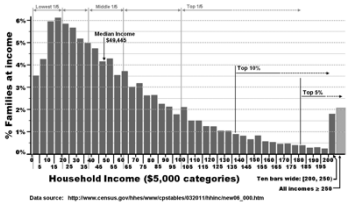


16

Example: Wine ratings (Whites)

Country	Mean = M = (ΣX) / N	Median
Argentina	71.20	71.00
Australia	86.81	86.68
France	85.90	86.00
USA	88.62	88.65

Measures of central tendency



18

Measures of central tendency

- Mode is the score that occurs most often
  - The peak of a histogram
  - The rating that occurred the most
    - For example, the Argentina white, Mode = 70 – 72

Measures of central tendency

- Mode can be used for nominal variables
  - For example, names
    - Female, USA Sophia
    - Male, USA James
    - Female, France Emma
    - Male, France Nathan

Measures of central tendency

- Mode can be used for nominal variables
  - For example, names
    - Female, Argentina Sofia
    - Male, Argentina Juan
    - Female, Australia Charlotte
    - Male, Australia Oliver

Segment summary

- Measures of central tendency
  - Mean
  - Median
  - Mode

22

END SEGMENT

23

Lecture 4 ~ Segment 2

Measures of variability

24

Variability

- A measure that describes the range and diversity of scores in a distribution
  - *Standard deviation* (SD): the average deviation from the mean in a distribution
  - *Variance* = SD^2

Variability

- *Variance* = SD^2

$$SD^2 = [\Sigma(X - M)^2] / N$$

Variance

- Variation is natural and observed in all species and that's good:
  - *On the Origin of Species (1859)*
  - *Variation Under Domestication (1868)*

27

Linsanity!

28

Jeremy Lin (10 games)

Points per game	(X-M)	$(X-M)^2$
28	5.3	28.09
26	3.3	10.89
10	-12.7	161.29
27	4.3	18.49
20	-2.7	7.29
38	15.3	234.09
23	0.3	0.09
28	5.3	28.09
25	2.3	5.29
2	-20.7	428.49
M = 227/10 = 22.7	M = 0/10 = 0	M = 922.1/10 = 92.21

Results

- M = Mean = 22.7
- SD = Standard Deviation = 9.6
- SD^2 = Variance = 92.21

Notation

- M = Mean
- SD = Standard Deviation
- SD^2 = Variance (also known as MS)
  - MS stands for Mean Squares
  - SS stands for Sum of Squares

Lin vs. Kobe

10 games, R output

```
> # Descriptive statistics for the variables in the dataframe called ppg
> describe(ppg)
       var  n mean    sd median trimmed  mad min max range  skew kurtosis   se
Lin      1 10 22.7 10.12   25.5   23.38 3.71   2  38    36 -0.67    -0.46 3.20
Bryant   2 10 26.4  7.46   27.0   27.25 5.93  10  36    26 -0.77    -0.19 2.36
```

33

9 games, R output

```
> # Descriptive statistics for the variables in the dataframe called ppg
> describe(ppg)
       var  n  mean   sd median trimmed  mad min max range  skew kurtosis   se
Lin      1  9 25.00 7.47     26   25.00 2.97  10  38    28 -0.33    -0.14 2.49
Bryant   2  9 26.67 7.86     27   26.67 7.41  10  36    26 -0.82    -0.36 2.62
```

34

Summary statistics: Review

- Important concepts
  - Central tendency (mean, median, mode)
  - Variability (standard deviation and variance)

Summary statistics: Review

- Summary statistics (formulae to know)
  - $M = (\Sigma X) / N$
  - $SD^2 = [\Sigma(X - M)^2] / N$
    - Used for descriptive statistics
  - $SD^2 = [\Sigma(X - M)^2] / (N - 1)$
    - Used for inferential statistics

END SEGMENT

37

END LECTURE 4

38

Statistics One

Lecture 5
Correlation

1

Three segments

- Overview
- Calculation of *r*
- Assumptions

2

Lecture 5 ~ Segment 1

Correlation: Overview

3

Correlation: Overview

- Important concepts & topics
  - What is a correlation?
  - What are they used for?
  - Scatterplots
  - CAUTION!
  - Types of correlations

4

Correlation: Overview

- Correlation
  - A statistical procedure used to measure and describe the relationship between two variables
  - Correlations can range between +1 and -1
    - +1 is a perfect positive correlation
    - 0 is no correlation (independence)
    - -1 is a perfect negative correlation

5

Correlation: Overview

- When two variables, let's call them X and Y, are correlated, then one variable can be used to predict the other variable
  - More precisely, a person's score on X can be used to predict his or her score on Y

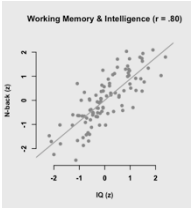
6

Correlation: Overview

- Example:
  - Working memory capacity is strongly correlated with intelligence, or IQ, in healthy young adults
  - So if we know a person's IQ then we can predict how they will do on a test of working memory

7

Correlation: Overview



8

Correlation: Overview

- CAUTION!
  - Correlation does not imply causation

9

Correlation: Overview

- CAUTION!
  - The magnitude of a correlation depends upon many factors, including:
    - Sampling (random and representative?)

10

Correlation: Overview

- CAUTION!
  - The magnitude of a correlation is also influenced by:
    - Measurement of X & Y (See Lecture 6)
    - Several other assumptions (See Segment 3)

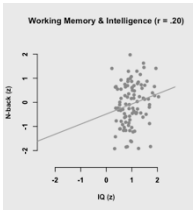
11

Correlation: Overview

- For now, consider just one assumption:
  - Random and representative sampling
  - There is a strong correlation between IQ and working memory among all healthy young adults.
    - What is the correlation between IQ and working memory among college graduates?

12

Correlation: Overview



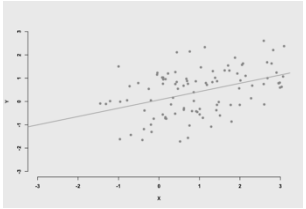
13

Correlation: Overview

- CAUTION!
- Finally & perhaps most important:
  - The correlation coefficient is a sample statistic, just like the mean
    - It may not be representative of ALL individuals
      - For example, in school I scored very high on Math and Science but below average on Language and History

14

Correlation: Overview



15

Correlation: Overview

- Note: there are several types of correlation coefficients, for different variable types
  - Pearson product-moment correlation coefficient (*r*)
    - When both variables, X & Y, are continuous
  - Point bi-serial correlation
    - When 1 variable is continuous and 1 is dichotomous

16

Correlation: Overview

- Note: there are several types of correlation coefficients
  - Phi coefficient
    - When both variables are dichotomous
  - Spearman rank correlation
    - When both variables are ordinal (ranked data)

17

Segment summary

- Important concepts/topics
  - What is a correlation?
  - What are they used for?
  - Scatterplots
  - CAUTION!
  - Types of correlations

18

END SEGMENT

19

Lecture 5 ~ Segment 2

Calculation of *r*

20

Calculation of *r*

- Important topics
  - *r*
    - Pearson product-moment correlation coefficient
      - Raw score formula
      - Z-score formula
  - Sum of cross products (SP) & Covariance

21

Calculation of *r*

- *r* = the degree to which X and Y vary together, relative to the degree to which X and Y vary independently
- *r* = (Covariance of X & Y) / (Variance of X & Y)

22

Calculation of *r*

- Two ways to calculate *r*
  - Raw score formula
  - Z-score formula

23

Calculation of *r*

- Let's quickly review calculations from Lecture 4 on summary statistics
  - Variance = SD^2 = MS = (SS/N)

24

Linsanity!

25

Jeremy Lin (10 games)

Points per game	(X-M)	$(X-M)^2$
28	5.3	28.09
26	3.3	10.89
10	-12.7	161.29
27	4.3	18.49
20	-2.7	7.29
38	15.3	234.09
23	0.3	0.09
28	5.3	28.09
25	2.3	5.29
2	-20.7	428.49
M = 227/10 = 22.7	M = 0/10 = 0	M = 922.1/10 = 92.21

26

Results

- M = Mean = 22.7
- SD^2 = Variance = MS = SS/N = 92.21
- SD = Standard Deviation = 9.6

27

Just one new concept!

- SP = Sum of cross Products

28

Just one new concept!

- Review: To calculate SS
  - For each row, calculate the deviation score
    - $(X - M_x)$
  - Square the deviation scores
    - $(X - M_x)^2$
  - Sum the squared deviation scores
    - $SS_x = \Sigma[(X - M_x)^2] = \Sigma[(X - M_x) \times (X - M_x)]$

29

Just one new concept!

- To calculate SP
  - For each row, calculate the deviation score on X
    - $(X - M_x)$
  - For each row, calculate the deviation score on Y
    - $(Y - M_y)$

30

Just one new concept!

- To calculate SP
  - Then, for each row, multiply the deviation score on X by the deviation score on Y
    - $(X - M_x) \times (Y - M_y)$
  - Then, sum the "cross products"
    - $SP = \Sigma[(X - M_x) \times (Y - M_y)]$

31

Calculation of *r*

Raw score formula:

$r = SP_{xy} / \mathbf{SQRT}(SS_x \times SS_y)$

32

Calculation of *r*

$SP_{xy} = \Sigma[(X - M_x) \times (Y - M_y)]$

$SS_x = \Sigma(X - M_x)^2 = \Sigma[(X - M_x) \times (X - M_x)]$

$SS_y = \Sigma(Y - M_y)^2 = \Sigma[(Y - M_y) \times (Y - M_y)]$

33

Formulae to calculate *r*

$r = SP_{xy}$ / **SQRT** $(SS_x \times SS_y)$

$r = \Sigma[(X - M_x) \times (Y - M_y)]$ /
SQRT $(\Sigma(X - M_x)^2 \times \Sigma(Y - M_y)^2)$

34

Formulae to calculate *r*

Z-score formula:

$r = \Sigma(Z_x \times Z_y) / N$

35

Formulae to calculate *r*

$Z_x = (X - M_x) / SD_x$
$Z_y = (Y - M_y) / SD_y$

$SD_x = SQRT\ (\Sigma(X - M_x)^2 / N)$
$SD_y = SQRT\ (\Sigma(Y - M_y)^2 / N)$

36

Formulae to calculate *r*

Proof of equivalence:

$Z_x = (X - M_x) / SQRT (\Sigma(X - M_x)^2 / N)$

$Z_y = (Y - M_y) / SQRT (\Sigma(Y - M_y)^2 / N)$

37

Formulae to calculate *r*

$r = \Sigma \{ [(X - M_x) / SQRT (\Sigma(X - M_x)^2 / N)] \times [(Y - M_y) / SQRT (\Sigma(Y - M_y)^2 / N)] \} / N$

38

Formulae to calculate *r*

$r = \Sigma \{ [(X - M_x) / SQRT (\Sigma(X - M_x)^2 / N)] \times [(Y - M_y) / SQRT (\Sigma(Y - M_y)^2 / N)] \} / N$

$r = \Sigma [(X - M_x) \times (Y - M_y)] / SQRT (\Sigma(X - M_x)^2 \times \Sigma(Y - M_y)^2)$

$r = SP_{xy} / SQRT (SS_x \times SS_y)$ ← The raw score formula!

39

Variance and covariance

- Variance = MS = SS / N
- Covariance = COV = SP / N

- Correlation is standardized COV
  - Standardized so the value is in the range -1 to 1

40

Note on the denominators

- Correlation for descriptive statistics
  - Divide by N
- Correlation for inferential statistics
  - Divide by N – 1

41

Segment summary

- Important topics
  - *r*
    - Pearson product-moment correlation coefficient
      - Raw score formula
      - Z-score formula
  - Sum of cross Products (SP) & Covariance

42

END SEGMENT

43

Lecture 5 ~ Segment 3

Assumptions

44

Assumptions

- Assumptions when interpreting *r*
  - Normal distributions for X and Y
  - Linear relationship between X and Y
  - Homoscedasticity

45

Assumptions

- Assumptions when interpreting *r*
  - Reliability of X and Y
  - Validity of X and Y
  - Random and representative sampling

46

Assumptions

- Assumptions when interpreting *r*
  - Normal distributions for X and Y
    - How to detect violations?
      - Plot histograms and examine summary statistics

47

Assumptions

- Assumptions when interpreting *r*
  - Linear relationship between X and Y
    - How to detect violation?
      - Examine scatterplots (see following examples)

48

Assumptions

- Assumptions when interpreting *r*
  - Homoscedasticity
    - How to detect violation?
      - Examine scatterplots (see following examples)

49

Homoscedasticity

- In a scatterplot the vertical distance between a dot and the regression line reflects the amount of prediction error (known as the "residual")

50

Homoscedasticity

- Homoscedasticity means that the distances (the residuals) are not related to the variable plotted on the X axis (they are not a function of X)
- This is best illustrated with scatterplots

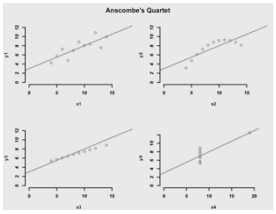
51

Anscombe's quartet

- In 1973, statistician Dr. Frank Anscombe developed a classic example to illustrate several of the assumptions underlying correlation and regression

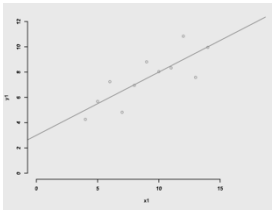
52

Anscombe's quartet



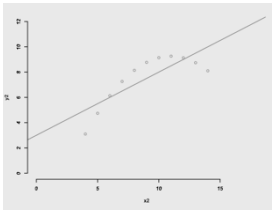
53

Anscombe's quartet



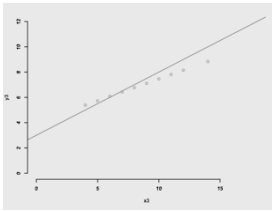
54

Anscombe's quartet



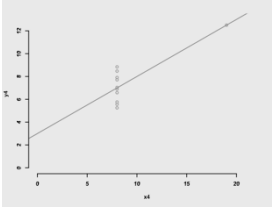
55

Anscombe's quartet



56

Anscombe's quartet



57

Segment summary

- Assumptions when interpreting *r*
  - Normal distributions for X and Y
  - Linear relationship between X and Y
  - Homoscedasticity

58

Segment summary

- Assumptions when interpreting *r*
  - Reliability of X and Y
  - Validity of X and Y
  - Random and representative sampling

59

END SEGMENT

60

END LECTURE 5

61

Statistics One

Lecture 6
Measurement

1

Three segments

- Reliability
- Validity
- Sampling

2

Lecture 6 ~ Segment 1

Reliability

3

Reliability

- Important concepts & topics
  - Classical test theory
  - Reliability estimates

4

Reliability

- Classical test theory
  - Raw scores (X) are not perfect
  - They are influenced by bias and chance error
    - For example, measurement of body temperature

5

Reliability

- Classical test theory
  - In a perfect world, it would be possible to obtain a "true score" rather than a "raw score" (X)
    - X = true score + bias + error
    - This is also known as "true score theory"

6

Reliability

- A measure (X) is considered to be reliable as it approaches the true score
  - The problem is we don't know the true score
  - So, we estimate reliability

7

Reliability

- Methods to estimate reliability
  - Test / re-test
  - Parallel tests
  - Inter-item estimates

8

Reliability

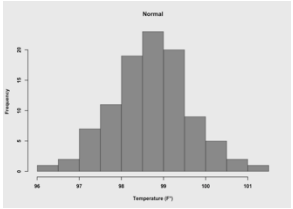
- Example: Body temperature
  - Orally
  - Internally
  - Infrared thermometer: “The wand”

9

Body temperature

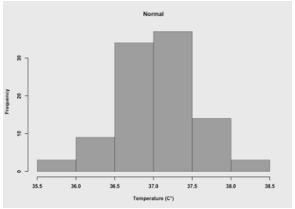
10

Body temperature F°



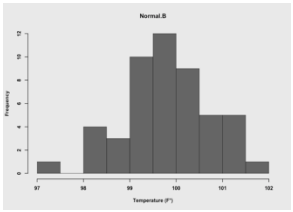
11

Body temperature C°



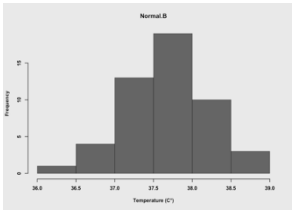
12

Body temperature F°: Biased



13

Body temperature C°: Biased



14

Reliability

- Test / re-test
  - Measure everyone twice
    - X1
    - X2

15

Reliability

- Test / re-test
  - The correlation between X1 and X2 is an estimate of reliability
    - However, if the bias is uniform then we won't detect it with the test / re-test method

16

Reliability

- Parallel tests
  - Measure body temperature with the wand (X1) and with an oral thermometer (X2)
  - The correlation between X1 and X2 is an estimate of reliability
    - AND, now the bias of the wand will be revealed

17

Reliability

- Inter-item
  - Inter-item is the most commonly used method in the social sciences
    - Test / re-test and parallel tests are time consuming
    - Inter-item is therefore more cost efficient

18

Reliability

- Inter-item
  - For example, suppose a 20-item survey is designed to measure extraversion
    - Randomly select 10 items to get sub-set A (X1)
    - The other 10 items become sub-set B (X2)
    - The correlation between X1 and X2 is an estimate of reliability

19

Segment summary

- Classical test theory (true score theory)
- Reliability estimates
  - Test / re-test
  - Parallel tests
  - Inter-item estimates

20

END SEGMENT

21

Lecture 6 ~ Segment 2

Validity

22

Validity

- What is a construct?
  - How to operationalize a construct
  - Construct validity
    - Content validity
    - Convergent validity
    - Divergent validity
    - Nomological validity

23

Validity

- What is a construct?
  - An ideal "object" that is not directly observable
    - As opposed to "real" observable objects
    - For example, "intelligence" is a construct

24

Validity

- How do we operationalize a construct?
  - The process of defining a construct to make it observable and quantifiable
    - For example, intelligence tests

25

Validity

- Construct validity
  - Content validity
  - Convergent validity
  - Divergent validity
  - Nomological validity

26

Validity

- An example:
  - Construct: Verbal ability in children

27

Validity

- How to operationalize?
  - A vocabulary test

28

Validity

- Construct validity
  - Content validity
    - Does the test consist of words that children in the population and sample should know?

29

Validity

- Construct validity
  - Convergent validity
    - Does the test correlate with other, established measures of verbal ability?
      - For example, reading comprehension

30

Validity

- Construct validity
  - Divergent validity
    - Does the test correlate less well with measures designed to test a different type of ability?
      - For example, spatial ability

31

Validity

- Construct validity
  - Nomological validity
    - Are scores on the test consistent with more general theories, for example, of child development and neuroscience
      - For example, a child with neural damage or disease to brain regions associated with language development should score lower on the test

32

Reliability & Validity: Review

- Important concepts & topics
  - Classical test theory
  - Reliability estimates
  - Construct validity

33

END SEGMENT

34

Lecture 6 ~ Segment 3

Sampling

35

Sampling

- Important concepts & topics
  - Random and representative sampling
  - Sampling error
  - Standard error

36

Sampling

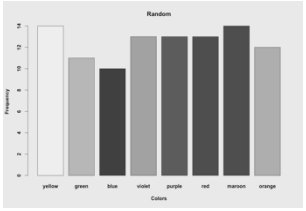
- Random and representative
- Recall the color wheel from Lecture 1

37

Illustration

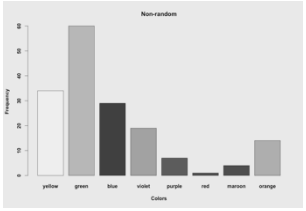
38

Random



39

Not random



40

Sampling error

- *Sampling error*: The difference between the population and the sample
  - Notice that even the "random" histogram is not "perfectly" random
  - There is some fluctuation due to *sampling error*

41

Sampling error

- PROBLEM!
  - We typically don't know the population parameters
  - So, how do we estimate sampling error?

42

Sampling error

- Sampling error mainly depends on the size of the sample, relative to the size of the population
  - As sample size increases, sampling error decreases
- It also depends on the variance in the population

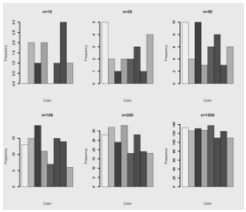
43

Sampling error

- Assume 6 samples from a normal population
  - N = 10
  - N = 20
  - N = 50
  - N = 100
  - N = 200
  - N = 1000

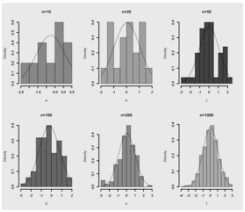
44

Sampling error



45

Sampling error



46

Sampling error

- Sampling error is estimated from the size of the sample and the variance in the sample
  - Under the assumption that the sample is random and representative of the population

47

Standard error

- Standard error is an estimate of amount of sampling error
  - SE = SD / SQRT(N)
  - SE: Standard error
  - SD: Standard deviation of the sample
  - N: Size of the sample

48

Segment Summary

- Important concepts & topics
  - Random and representative sampling
  - Sampling error
  - Standard error

49

END SEGMENT

50

END LECTURE 6

51

Statistics One

Lecture 7
Introduction to Regression

1

Three segments

- Overview
- Calculation of regression coefficients
- Assumptions

2

Lecture 7 ~ Segment 1

Regression: Overview

3

Regression: Overview

- Important concepts & topics
  - Simple regression vs. multiple regression
  - Regression equation
  - Regression model

4

Regression: Overview

- *Regression*: a statistical analysis used to predict scores on an outcome variable, based on scores on one or multiple predictor variables
  - *Simple regression*: one predictor variable
  - *Multiple regression*: multiple predictors

5

Regression: Overview

- Example: IMPACT (see Lab 2)
  - An online assessment tool to investigate the effects of sports-related concussion
    - http://www.impacttest.com

6

IMPACT example

- IMPACT provides data on 6 variables
  - Verbal memory
  - Visual memory
  - Visual motor speed
  - Reaction time
  - Impulse control
  - Symptom score

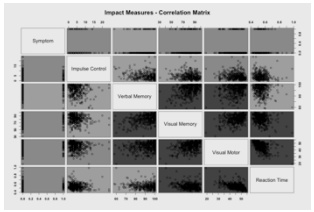
7

IMPACT: Correlations pre-injury

```
> cor(impact)
                Verbal Memory Visual Memory Visual Motor Reaction Time Impulse Control     Symptom
Verbal Memory      1.00000000    0.41549808   0.24573123   -0.15638810     -0.18184017 -0.09333058
Visual Memory      0.41549808    1.00000000   0.34044313   -0.25796852     -0.10059464 -0.06243145
Visual Motor       0.24573123    0.34044313   1.00000000   -0.50452903     -0.07151656 -0.09090637
Reaction Time     -0.15638810   -0.25796852  -0.50452903    1.00000000     -0.10547302  0.02403135
Impulse Control   -0.18184017   -0.10059464  -0.07151656   -0.10547302      1.00000000  0.02908636
Symptom           -0.09333058   -0.06243145  -0.09090637    0.02403135      0.02908636  1.00000000
```

8

IMPACT: Correlations pre-injury



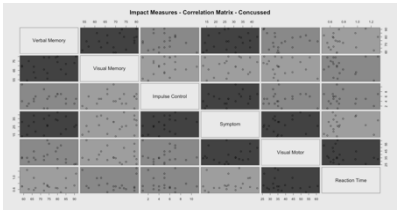
9

IMPACT: Correlations post-injury

	Verbal_Memory	Visual_Memory	Visual_Motor	Reaction_Time	Impulse_Control	Symptom
Verbal_Memory	1.00000000	0.3469979	-0.059072912	0.1205644	-0.050532262	0.2161228
Visual_Memory	0.34699791	1.0000000	-0.155630881	-0.0975997	0.183486515	0.2067446
Visual_Motor	-0.05907291	-0.1556309	1.000000000	-0.2863136	0.004794629	0.2281889
Reaction_Time	0.12056437	-0.0975997	-0.286313634	1.0000000	-0.042379705	0.1477275
Impulse_Control	-0.05053226	0.1834865	0.004794629	-0.0423797	1.000000000	0.4008124
Symptom	0.21612276	0.2067446	0.228188865	0.1477275	0.400812421	1.0000000

10

IMPACT: Correlations post-injury



11

IMPACT example

- For this example, assume:
  - Symptom Score is the outcome variable
  - *Simple regression* example:
    - Predict Symptom Score from just one variable
  - *Multiple regression* example:
    - Predict Symptom Score from two variables

12

Regression equation

- $Y = m + bX + e$
  - Y is a linear function of X
  - m = intercept
  - b = slope
  - e = error (residual)

13

Regression equation

- $Y = B_0 + B_1X_1 + e$
  - Y is a linear function of X_1
  - B_0 = intercept = regression constant
  - B_1 = slope = regression coefficient
  - e = error (residual)

14

Model R and R^2

- R = multiple correlation coefficient
  - $R = r_{\hat{Y}Y}$
  - The correlation between the predicted scores and the observed scores
- R^2
  - The percentage of variance in Y explained by the model

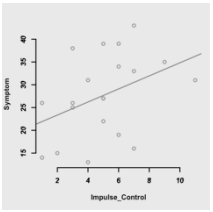
15

IMPACT example

- $Y = B_0 + B_1X_1 + e$
  - Let Y = Symptom Score
  - Let X_1 = Impulse Control
  - Solve for B_0 and B_1
  - In R, function lm

16

IMPACT example



$\hat{Y} = 20.48 + 1.43(X)$

$r = .40$

$R^2 = 16\%$

17

IMPACT example

```
> model1 <- lm(Symptom ~ Impulse_Control)
> summary(model1)

Call:
lm(formula = Symptom ~ Impulse_Control)

Residuals:
    Min      1Q  Median      3Q     Max
-14.5189  -6.2156   0.7189   4.8172  13.2189

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.4779     4.3098   4.752 0.000159 ***
Impulse_Control  1.4344     0.7728   1.856 0.079884 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.536 on 18 degrees of freedom
Multiple R-squared:  0.1607, Adjusted R-squared:  0.114
F-statistic: 3.445 on 1 and 18 DF,  p-value: 0.07988
```

18

Regression model

- The regression model is used to model or predict future behavior
  - The model is just the regression equation
  - Later in the course we will discuss more complex models that consist of a set of regression equations

19

Regression: It gets better

- The goal is to produce better models so we can generate more accurate predictions
  - Add more predictor variables, and/or…
  - Develop better predictor variables

20

IMPACT example

- $Y = B_0 + B_1X_1 + B_2X_2 + e$
  - Let Y = Symptom Score
  - Let X_1 = Impulse Control
  - Let X_2 = Verbal Memory
  - Solve for B_0 and B_1 and B_2
  - In R, function lm

21

IMPACT example

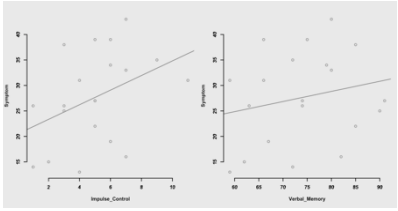
```
> model2 <- lm(Symptom ~ Impulse_Control + Verbal_Memory)
> summary(model2)

Call:
lm(formula = Symptom ~ Impulse_Control + Verbal_Memory)

Residuals:
    Min      1Q  Median      3Q     Max
-16.337  -6.018   0.230   4.848  13.104

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.1313    15.4322   0.268   0.7921
Impulse_Control  1.4773     0.7692   1.921   0.0717
Verbal_Memory    0.2179     0.1976   1.103   0.2856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.485 on 17 degrees of freedom
Multiple R-squared:  0.2167, Adjusted R-squared:  0.1245
F-statistic: 2.351 on 2 and 17 DF,  p-value: 0.1255
```

$\hat{Y} = 4.13 + 1.48(X_1) + 0.22(X_2)$

$R^2 = 22\%$

22

IMPACT example



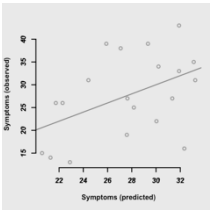
23

Model R and R^2

- R = multiple correlation coefficient
  - $R = r_{\hat{Y}Y}$
  - The correlation between the predicted scores and the observed scores
- R^2
  - The percentage of variance in Y explained by the model

24

IMPACT example



$R^2 = 22\%$

$r_{\hat{Y}Y} = .47$

25

Segment summary

- Important concepts & topics
  - Simple regression vs. multiple regression
  - Regression equation
  - Regression model

26

END SEGMENT

27

Lecture 7 ~ Segment 2

Calculation of regression coefficients

28

Estimation of coefficients

- Regression equation:
  - $Y = B_0 + B_1X_1 + e$
  - $\hat{Y} = B_0 + B_1X_1$
    - $(Y - \hat{Y}) = e$ (residual)

29

Estimation of coefficients

- The values of the coefficients (e.g., B_1) are estimated such that the regression model yields optimal predictions
  - Minimize the residuals!

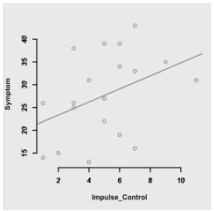
30

Estimation of coefficients

- *Ordinary Least Squares* estimation
  - Minimize the sum of the squared (SS) residuals
  - $SS.RESIDUAL = \Sigma(Y - \hat{Y})^2$

31

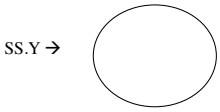
IMPACT example



32

Estimation of coefficients

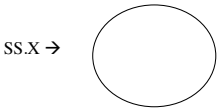
- Sum of Squared deviation scores (SS) in variable Y
  - SS.Y



33

Estimation of coefficients

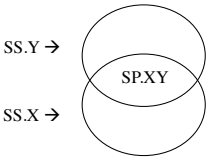
- Sum of Squared deviation scores (SS) in variable X
  - SS.X



34

Estimation of coefficients

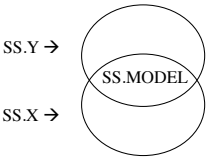
- Sum of Cross Products
  - SP.XY



35

Estimation of coefficients

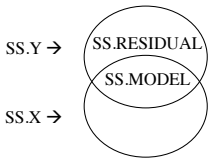
- Sum of Cross Products = SS of the Model
  - SP.XY = SS.MODEL



36

Estimation of coefficients

- SS.RESIDUAL = (SS.Y – SS.MODEL)



37

Estimation of coefficients

- Formula for the unstandardized coefficient
  - $B_1 = r \times (SD_y / SD_x)$

38

Estimation of coefficients

- Formula for the standardized coefficient
  - If X and Y are standardized then
    - $SD_y = SD_x = 1$
    - $B = r \times (SD_y / SD_x)$
    - $\beta = r$

39

Segment summary

- Important concepts
  - Regression equation and model
  - Ordinary least squares estimation
  - Unstandardized regression coefficients
  - Standardized regression coefficients

40

END SEGMENT

41

Lecture 7 ~ Segment 3

Assumptions

42

Assumptions

- Assumptions of linear regression
  - Normal distribution for Y
  - Linear relationship between X and Y
  - Homoscedasticity

43

Assumptions

- Assumptions of linear regression
  - Reliability of X and Y
  - Validity of X and Y
  - Random and representative sampling

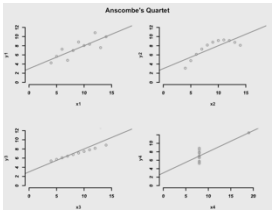
44

Assumptions

- Assumptions of linear regression
  - Normal distribution for Y
  - Linear relationship between X and Y
  - Homoscedasticity

45

Anscombe's quartet



46

Anscombe's quartet

- Regression equation for all 4 examples:
  - $\hat{Y} = 3.00 + 0.50(X_1)$

47

Anscombe's quartet

- To test assumptions, save residuals
  - $Y = B_0 + B_1X_1 + e$
  - $e = (Y - \hat{Y})$

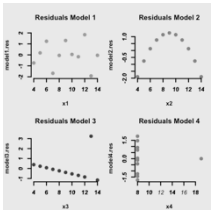
48

Anscombe's quartet

- Then examine a scatterplot with
  - X on the X-axis
  - Residuals on the Y-axis

49

Anscombe's quartet



50

Segment summary

- Assumptions when interpreting *r*
  - Normal distributions for Y
  - Linear relationship between X and Y
  - Homoscedasticity
  - Examine residuals to evaluate assumptions

51

END SEGMENT

52

END LECTURE 7

53

Statistics One

Lecture 8
Null Hypothesis Significance Testing
(NHST)

1

Two segments

- Overview
- Problems & Remedies

2

Lecture 8 ~ Segment 1

NHST: Overview

3

NHST: Overview

- Null Hypothesis Significance Testing (NHST)
  - H_0 = null hypothesis
    - For example, $r = 0$
  - H_A = alternative hypothesis
    - For example, $r > 0$

4

NHST: Overview

- Null Hypothesis Significance Testing (NHST)
  - H_0 = null hypothesis
    - For example, B = 0
  - H_A = alternative hypothesis
    - For example, B > 0

5

NHST: Overview

- If the alternative hypothesis predicts the direction of the relationship between X & Y (positive vs. negative)
  - Directional test
  - Otherwise it is known as a non-directional test

6

NHST: Overview

- Null Hypothesis Significance Testing (NHST)
  - H_0 = null hypothesis
    - For example, B = 0
  - H_A = alternative hypothesis
    - For example, B != 0

7

NHST: Overview

- Assume H_0 is true, then calculate the probability of observing data with these characteristics, given that H_0 is true
  - $p = P(D \mid H_0)$
  - If p is very low, then Reject H_0, else Retain H_0

8

NHST: Overview

Experimenter Decision

Truth	Retain H_0	Reject H_0
H_0 true	Correct Decision	Type I error (False alarm)
H_0 false	Type II error (Miss)	Correct Decision

9

NHST: Overview

- p = P(D | H_0)
- Given that the null hypothesis is true, the probability of these, or more extreme data, is p
  - NOT: The probability of the null hypothesis being true is p
  - In other words, P(D|H_0) != P(H_0|D)

10

NHST so far in this course

- *r*
  - Is the correlation significantly different from zero?
- B
  - Is the slope of the regression line for X significantly different from zero?

11

NHST for B

- t = B / SE
  - B is the unstandardized regression coefficient
  - SE = standard error
  - SE = SQRT[SS.RESIDUAL / (N – 2)]

12

Segment summary

- NHST is a procedure for hypothesis testing
- Requires a binary decision
  - Reject or Retain the Null Hypothesis
- Four possible outcomes
  - Correct retention, correction rejection
  - False alarm (Type 1 error), Miss (Type II error)

13

Lecture 8 ~ Segment 2

NHST Problems & Remedies

14

NHST Problems

- Biased by sample size
- Arbitrary decision rule
- Yokel local test
- Error prone
- Shady logic

15

NHST Problems

- Biased by sample size
  - For example, in regression
    - p-value is based on t-value
    - t = B / SE
    - SE = SQRT(SS.RESIDUAL / (N – 2))

16

NHST Problems

- Arbitrary decision rule
  - The cut-off value (alpha) is arbitrary
  - p < .05 is considered standard but still arbitrary
  - Problems arise when p is close to .05 but not less than .05

17

NHST Problems

- Yokel local test
  - Many researchers use NHST because it's the only approach they know
  - NHST encourages weak hypothesis testing

18

NHST Problems

- Error prone
  - Type I errors
    - Probability of Type I errors increases when researchers conduct multiple NHSTs
  - Type II errors
    - Many fields of research are plagued by a large degree of sampling error, which makes it difficult to detect an effect, even when the effect exists

19

NHST Problems

- Shady logic
- Modus tollens
  - If p then q
  - Not q
  - Therefore, not p
  - If the null hypothesis is correct, then these data can not occur
  - The data have occurred
  - Therefore, the null hypothesis is false

20

NHST Problems

- Shady logic
- If the null hypothesis is correct, then these data are highly unlikely
- These data have occurred
- Therefore, the null hypothesis is highly unlikely
- If a person plays football, then he or she is probably not a professional player
- This person is a professional player
- Therefore, he or she probably does not play football

21

NHST Remedies

- Biased by sample size
  - Supplement all NHSTs with estimates of effect size
    - For example, in regression, report standardized regression coefficients and the model R-squared

22

NHST Remedies

- Arbitrary decision rule
  - Again, supplement NHST with estimates of effect size
  - Also, avoid phrases such as “marginally significant” or “highly significant”

23

NHST Remedies

- Yokel local test
  - Learn other forms of hypothesis testing
  - Consider multiple alternative hypotheses
    - Model comparison

24

NHST Remedies

- Error prone
  - Replicate significant effects to avoid long-term impact of Type 1 errors
  - Obtain large and representative samples to avoid Type II errors

25

NHST Remedies

- Shady logic
  - Simply remember, $p = P(D \mid H_0)$
  - OR, avoid NHST, and…
  - Report Confidence Intervals only (see Lecture 10)
  - Apply Bayesian inference

26

NHST Problems

- Biased by sample size
- Arbitrary decision rule
- Yokel local test
- Error prone
- Shady logic

27

END SEGMENT

28

END LECTURE 8

29

Statistics One

Lecture 9
The Central Limit Theorem

1

Two segments

- Sampling distributions
- Central limit theorem

2

Lecture 9 ~ Segment 1

Sampling distributions

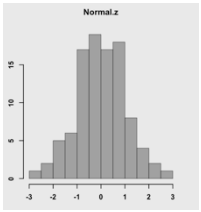
3

Review of histograms

- Histograms are used to display distributions
- For example, the body temperature of a random sample of healthy people

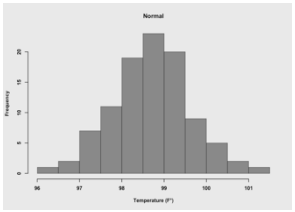
4

Review of histograms



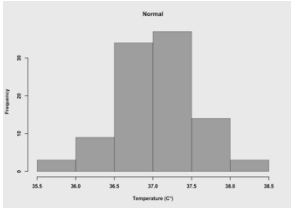
5

Review of histograms



6

Review of histograms



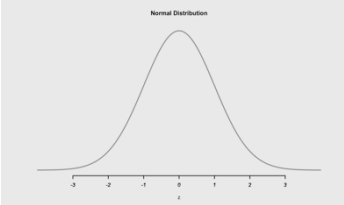
7

Review of histograms

- If a distribution is perfectly normal then the properties of the distribution are known

8

The normal distribution



9

The normal distribution & probability

- This allows for predictions about the distribution
  - Predictions aren't certain
  - They are *probabilistic*

10

The normal distribution & probability

- If one person is randomly selected from the sample, what is the probability that his or her body temperature is less than Z = 0?
  - Easy, p = .50

11

The normal distribution & probability

- If one person is randomly selected from the sample, what is the probability that his or her body temperature is greater than Z = 2? (100 F°, 38 C°)?
  - p = .02

12

The normal distribution & probability

- If this sample is healthy, then no one should have a fever
- I detected a person with a fever
- Therefore, this sample is not 100% healthy

13

Sampling distribution

- A distribution of sample statistics, obtained from multiple samples
  - For example,
    - Distribution of sample means
    - Distribution of sample correlations
    - Distribution of sample regression coefficients

14

Sampling distribution

- It is hypothetical
  - Assume a mean is calculated from a sample, obtained randomly from the population
  - Assume a certain sample size, N
  - Now, assume we had multiple random samples, all of size N, and therefore many sample means
  - Collectively, they form a *sampling distribution*

15

Sampling distribution & probability

- If one sample is obtained from a normal healthy population, what is the probability that the sample mean is less than Z = 0?
  - Easy, p = .50

16

Sampling distribution & probability

- If one sample is obtained from a normal healthy population, what is the probability that the sample mean is greater than Z = 2 (100 F°, 38 C°)?
  - p = .02

17

Sampling distribution & probability

- If this population is healthy, then no one sample should have a high mean body temperature
- I obtained a very high sample mean
- Therefore, the population is not healthy

18

Sampling distribution

- A distribution of sample statistics, obtained from multiple samples, each of size N
  - Distribution of sample means
  - Distribution of sample correlations
  - Distribution of sample regression coefficients

19

END SEGMENT

20

Lecture 9 ~ Segment 2

The Central Limit Theorem

21

Central Limit Theorem

- Three principles
  - The mean of a sampling distribution is the same as the mean of the population
  - The standard deviation of the sampling distribution is the square root of the variance of sampling distribution $\sigma^2 = \sigma^2 /N$
  - The shape of a sampling distribution is approximately normal if either (a) N >= 30 or (b) the shape of the population distribution is normal

22

NHST & Central limit theorem

- Multiple regression
  - Assume the null hypothesis is true
  - Conduct a study
  - Calculate B, SE, and t
  - t = B/SE

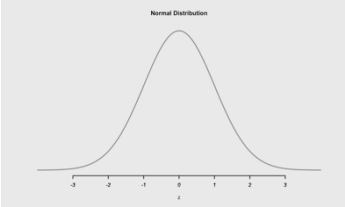
23

NHST & Central limit theorem

- Multiple regression
  - If the null hypothesis is true (B=0), then no one sample should have a very low or very high B
  - I obtained a very high B
  - Therefore, Reject the null hypothesis

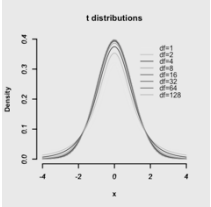
24

The normal distribution



25

The family of t distributions



26

NHST & Central limit theorem

- Multiple regression
  - Assume the null hypothesis is true
  - Conduct a study
  - Calculate B, SE, and t
  - t = B/SE
  - p-value is a function of t and sample size

27

NHST & the central limit theorem

- Multiple regression
  - If the null hypothesis is true (B=0), then no one sample should have a very low or very high B
  - I obtained a very high B
  - Therefore, Reject the null hypothesis
  - Very high and very low is $p < .05$

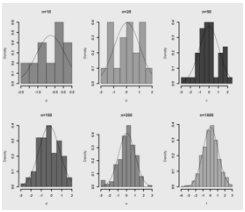
28

NHST & the central limit theorem

- Remember that sampling error, and therefore standard error, is largely determined by sample size

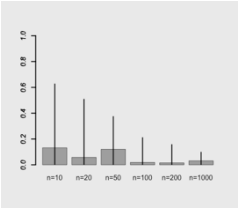
29

Sampling error and sample size



30

Sampling error and sample size



31

Central Limit Theorem

- Three principles
  - The mean of a sampling distribution is the same as the mean of the population
  - The standard deviation of the sampling distribution is the square root of the variance of sampling distribution $\sigma^2 = \sigma^2 /N$
  - The shape of a sampling distribution is approximately normal if either (a) N >= 30 or (b) the shape of the population distribution is normal

32

END SEGMENT

33

END LECTURE 9

34

Statistics One

Lecture 10
Confidence intervals

1

Two segments

- Confidence intervals for sample means (M)
- Confidence intervals for regression coefficients (B)

2

Lecture 10 ~ Segment 1

Confidence intervals for sample means (M)

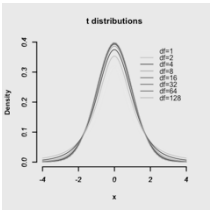
3

Confidence intervals

- All sample statistics, for example, a sample mean (M), are *point estimates*
- More specifically, a sample mean (M) represents a single point in a sampling distribution

4

The family of t distributions



5

Confidence intervals

- The logic of confidence intervals is to report a range of values, rather than a single value
- In other words, report an *interval estimate* rather than a *point estimate*

6

Confidence intervals

- *Confidence interval*: an interval estimate of a population parameter, based on a random sample
  - Degree of confidence, for example 95%, represents the probability that the interval captures the true population parameter

7

Confidence intervals

- The main argument for interval estimates is the reality of sampling error
- Sampling error implies that point estimates will vary from one study to the next
- A researcher will therefore be more confident about accuracy with an interval estimate

8

Confidence intervals for M

- Example, IMPACT
- Assume N = 30 and multiple samples…
  - Symptom Score (Baseline), M = 0.05
  - Symptom Score (Baseline), M = 0.07
  - Symptom Score (Baseline), M = 0.03
  - …

9

Confidence intervals for M

- Example, IMPACT
- Assume N = 10 and multiple samples…
  - Symptom Score (Baseline), M = 0.01
  - Symptom Score (Baseline), M = 0.20
  - Symptom Score (Baseline), M = 0.00
  - …

10

Confidence intervals for M

- Example, IMPACT
- Assume N = 30 and multiple samples…
  - Symptom Score (Post-injury), M = 12.03
  - Symptom Score (Post-injury), M = 12.90
  - Symptom Score (Post-injury), M = 14.13
  - …

11

Confidence intervals for M

- Example, IMPACT
- Assume N = 10 and multiple samples…
  - Symptom Score (Post-injury), M = 19.70
  - Symptom Score (Post-injury), M = 8.40
  - Symptom Score (Post-injury), M = 13.30
  - …

12

Confidence intervals for M

- The width of a confidence interval is influenced by
  - Sample size
  - Variance in the population (and sample)
  - Standard error (SE) = SD / SQRT(N)

13

Confidence intervals for M

- Example, IMPACT
- Assume N = 30
  - Symptom Score (Baseline)
  - 95% confidence interval
    - -0.03 – 0.10

14

Confidence intervals for M

- Example, IMPACT
- Assume N = 10
  - Symptom Score (Baseline)
  - 95% confidence interval
    - -0.10 – 0.50

15

Confidence intervals for M

- Example, IMPACT
- Assume N = 30
  - Symptom Score (Post-injury)
  - 95% confidence interval
    - 7.5 – 18.3

16

Confidence intervals for M

- Example, IMPACT
- Assume N = 10
  - Symptom Score (Post-injury)
  - 95% confidence interval
    - 2.7 – 23.9

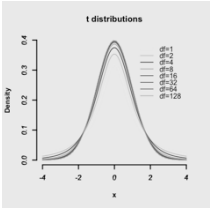
17

Confidence interval for M

- Upper bound = M + t(SE)
- Lower bound = M – t(SE)
- SE = SD / SQRT(N)
- t depends on level of confidence desired and sample size

18

The family of t distributions



19

Segment summary

- All sample statistics, for example, a sample mean (M), are *point estimates*
- More specifically, a sample mean (M) represents a single point in a sampling distribution

20

Segment summary

- The logic of confidence intervals is to report a range of values, rather than a single value
- In other words, report an *interval estimate* rather than a *point estimate*

21

Segment summary

- The width of a confidence interval is influenced by
  - Sample size
  - Variance in the population (and sample)
  - Standard error (SE) = SD / SQRT(N)

22

END SEGMENT

23

Lecture 10 ~ Segment 2

Confidence intervals for regression coefficients (B)

24

Confidence intervals for B

- All sample statistics, for example, a regression coefficient (B), are *point estimates*
- More specifically, a regression coefficient (B) represents a single point in a sampling distribution

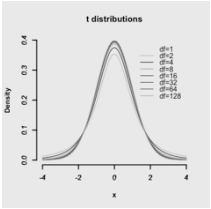
25

Confidence intervals for B

- In regression, t = B / SE

26

The family of t distributions



27

Confidence intervals for B

- The logic of confidence intervals is to report a range of values, rather than a single value
- In other words, report an *interval estimate* rather than a *point estimate*

28

Confidence intervals for B

- The main argument for interval estimates is the reality of sampling error
- Sampling error implies that point estimates will vary from one study to the next
- A researcher will therefore be more confident about accuracy with an interval estimate

29

Confidence intervals for B

- The width of a confidence interval is influenced by
  - Sample size
  - Variance in the population (and sample)
  - Standard error (SE) = SD / SQRT(N)

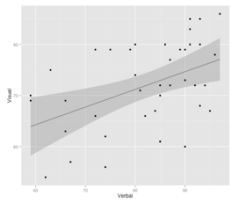
30

Confidence intervals for B

- Example, IMPACT
- Assume N = 40
  - Visual memory = B_0 + (B)Verbal memory

31

Confidence intervals for B

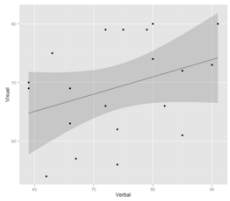


Confidence intervals for B

- Example, IMPACT
- Assume N = 20
  - Visual memory = B_0 + (B)Verbal memory

33

Confidence intervals for B



34

Segment summary

- All sample statistics are *point estimates*
- More specifically, a sample mean (M) or a regression coefficient (B) represents a single point in a sampling distribution

35

Segment summary

- The logic of confidence intervals is to report a range of values, rather than a single value
- In other words, report an *interval estimate* rather than a *point estimate*

36

Segment summary

- The width of a confidence interval is influenced by
  - Sample size
  - Variance in the population (and sample)
  - Standard error (SE) = SD / SQRT(N)

37

END SEGMENT

38

END LECTURE 10

39

Statistics One

Lecture 11
Multiple Regression

1

Three segments

- Multiple Regression (MR)
- Matrix algebra
- Estimation of coefficients

2

Lecture 11 ~ Segment 1

Multiple Regression

3

Multiple Regression

- Important concepts/topics
  - Multiple regression equation
  - Interpretation of regression coefficients

4

Simple vs. multiple regression

- Simple regression
  - Just one predictor (X)
- Multiple regression
  - Multiple predictors (X1, X2, X3, …)

5

Multiple regression

- Multiple regression equation
  - Just add more predictors (multiple Xs)

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + \ldots + B_kX_k$$

$$\hat{Y} = B_0 + \Sigma(B_kX_k)$$

6

Multiple regression

- Multiple regression equation

$\hat{Y}$ = predicted value on the outcome variable Y
B_0 = predicted value on Y when all X = 0
X_k = predictor variables
B_k = unstandardized regression coefficients
$Y - \hat{Y}$ = residual (prediction error)
k = the number of predictor variables

7

Model R and R^2

- R = multiple correlation coefficient
  - $R = r_{\hat{Y}Y}$
  - The correlation between the predicted scores and the observed scores
- R^2
  - The percentage of variance in Y explained by the model

8

Multiple regression: Example

- Outcome measure (Y)
  - Faculty salary (Y)
- Predictors (X1, X2, X3)
  - Time since PhD (X1)
  - Number of publications (X2)
  - Gender (X3)

9

Summary statistics

	M	SD
Salary	$64,115	$17,110
Time	8.09	5.24
Publications	15.49	7.51

10

Multiple regression: Example

- Gender
  - Male = 0
  - Female = 1

11

Multiple regression: Example

- lm(Salary ~ Time + Pubs + Gender)
- $\hat{Y}$ = 46,911 + 1,382(Time) + 502(Pubs) + -3,484(Gender)

12

Table of coefficients

	B	SE	t	β	p
B_0	46,911				
Time	1,382	236	5.86	.42	< .01
Pubs	502	164	3.05	.22	< .01
Gender	-3,484	2,439	-1.43	-.10	.16

$\hat{Y}$ = 46,911 + 1,382(Time) + 502(Pubs) + -3,484(Gender)

13

Multiple regression: Example

- What is $46,911?
- What is $502?
- Who makes more money, men or women?
- According to this model, is the gender difference statistically significant?
- What is the strongest predictor of salary?

14

Multiple regression: Example

- $46,911 is the predicted salary for a male professor who just graduated and has no publications (predicted score when all X=0)
- $502 is the predicted change in salary associated with an increase of one publication, for professors who have been out of school for an average amount of time, averaged across men and women

15

Multiple regression: Example

- Who makes more money, men or women?
  - Trick question: Based on the output we can't answer this question
- According to this model, is the gender difference statistically significant?
  - No

16

Multiple regression: Example

- What is the strongest predictor of salary?
  - Time

17

Segment summary

- Important concepts/topics
  - Multiple regression equation
  - Interpretation of regression coefficients

18

END SEGMENT

19

Lecture 11 ~ Segment 2

Matrix algebra

20

Matrix algebra

- Important concepts/topics
  - Matrix addition/subtraction/multiplication
  - Special types of matrices
    - Correlation matrix
    - Sum of squares / Sum of cross products matrix
    - Variance / Covariance matrix

21

Matrix algebra

- A matrix is a rectangular table of known or unknown numbers, for example,

$$M = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 3 & 4 \\ 4 & 2 \end{pmatrix}$$

22

Matrix algebra

- The size, or order, of a matrix is given by identifying the number of rows and columns. The order of matrix M is 4x2

$$M = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 3 & 4 \\ 4 & 2 \end{pmatrix}$$

23

Matrix algebra

- The transpose of a matrix is formed by rewriting its rows as columns

$$M = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 3 & 4 \\ 4 & 2 \end{pmatrix} \qquad M^T = \begin{pmatrix} 1 & 5 & 3 & 4 \\ 2 & 1 & 4 & 2 \end{pmatrix}$$

24

Matrix algebra

- Two matrices may be added or subtracted only if they are of the same order

$$N = \begin{pmatrix} 2 & 3 \\ 4 & 5 \\ 1 & 2 \\ 3 & 1 \end{pmatrix}$$

25

Matrix algebra

- Two matrices may be added or subtracted only if they are of the same order

$$M + N = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 3 & 4 \\ 4 & 2 \end{pmatrix} + \begin{pmatrix} 2 & 3 \\ 4 & 5 \\ 1 & 2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 5 \\ 9 & 6 \\ 4 & 6 \\ 7 & 3 \end{pmatrix}$$

26

Matrix algebra

- Two matrices may be multiplied when the number of columns in the first matrix is equal to the number of rows in the second matrix.
- If so, then we say they are *conformable* for matrix multiplication.

27

Matrix algebra

- Matrix multiplication:

$$R = M^T * N \qquad R_{ij} = \sum (M^T_{ik} * N_{kj})$$

28

Matrix algebra

$$R = M^T * N = \begin{pmatrix} 1 & 5 & 3 & 4 \\ 2 & 1 & 4 & 2 \end{pmatrix} * \begin{pmatrix} 2 & 3 \\ 4 & 5 \\ 1 & 2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 37 & 38 \\ 18 & 21 \end{pmatrix}$$

29

Matrix algebra

- In the next ten slides we will go from a raw dataframe to a correlation matrix!

30

Raw dataframe

$$X_{np} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix}$$

31

Row vector of sums

$$T_{1p} = 1_{1n} * X_{np} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} * \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix} = \begin{pmatrix} 34 & 35 & 34 \end{pmatrix}$$

32

Row vector of means

$$M_{1p} = T_{1p} * N^{-1} = \begin{pmatrix} 34 & 35 & 34 \end{pmatrix} * 10^{-1} = \begin{pmatrix} 3.4 & 3.5 & 3.4 \end{pmatrix}$$

33

Matrix of means

$$M_{np} = 1_{n1} * M_{1p} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} * \begin{pmatrix} 3.4 & 3.5 & 3.4 \end{pmatrix} = \begin{pmatrix} 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \end{pmatrix}$$

34

Matrix of deviation scores

$$D_{np} = X_{np} - M_{np} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix} - \begin{pmatrix} 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \end{pmatrix} = \begin{pmatrix} -.4 & -1.5 & -.4 \\ -.4 & -1.5 & -.4 \\ -1.4 & .5 & .6 \\ .6 & -.5 & .6 \\ .6 & .5 & -.4 \\ 1.6 & .5 & -.4 \\ -1.4 & 1.5 & .6 \\ -.4 & -.5 & -1.4 \\ 1.6 & -.5 & .6 \\ -.4 & 1.5 & .6 \end{pmatrix}$$

35

SS / SP matrix

$$S_{xx} = D_{pn}{}^{T} * D_{np} = \begin{pmatrix} -.4 & -.4 & -1.4 & .6 & .6 & 1.6 & -1.4 & -.4 & 1.6 & -.4 \\ -1.5 & -1.5 & .5 & -.5 & .5 & .5 & 1.5 & -.5 & -.5 & 1.5 \\ -.4 & -.4 & .6 & .6 & -.4 & -.4 & .6 & -1.4 & .6 & .6 \end{pmatrix} * \begin{pmatrix} -.4 & -1.5 & -.4 \\ -.4 & -1.5 & -.4 \\ -1.4 & .5 & .6 \\ .6 & -.5 & .6 \\ .6 & .5 & -.4 \\ 1.6 & .5 & -.4 \\ -1.4 & 1.5 & .6 \\ -.4 & -.5 & -1.4 \\ 1.6 & -.5 & .6 \\ -.4 & 1.5 & .6 \end{pmatrix}$$

$$= \begin{pmatrix} 10.4 & -2.0 & -.6 \\ -2.0 & 10.5 & 3.0 \\ -.6 & 3.0 & 4.4 \end{pmatrix}$$

Variance / Covariance matrix

$$C_{xx} = S_{xx} * N^{-1} = \begin{pmatrix} 10.4 & -2.0 & -.6 \\ -2.0 & 10.5 & 3.0 \\ -.6 & 3.0 & 4.4 \end{pmatrix} * 10^{-1} = \begin{pmatrix} 1.04 & -.20 & -.06 \\ -.20 & 1.05 & .30 \\ -.06 & .30 & .44 \end{pmatrix}$$

37

SD matrix

$$S_{xx} = (\text{Diag}(C_{xx}))^{1/2} = \begin{pmatrix} 1.02 & 0 & 0 \\ 0 & 1.02 & 0 \\ 0 & 0 & .66 \end{pmatrix}$$

38

Correlation matrix

$$R_{xx} = S_{xx}^{-1} * C_{xx} * S_{xx}^{-1} =$$

$$\begin{pmatrix} 1.02^{-1} & 0 & 0 \\ 0 & 1.02^{-1} & 0 \\ 0 & 0 & .66^{-1} \end{pmatrix} * \begin{pmatrix} 1.04 & -.20 & -.06 \\ -.20 & 1.05 & .30 \\ -.06 & .30 & .44 \end{pmatrix} * \begin{pmatrix} 1.02^{-1} & 0 & 0 \\ 0 & 1.02^{-1} & 0 \\ 0 & 0 & .66^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} 1.00 & -.19 & -.09 \\ -.19 & 1.00 & .44 \\ -.09 & .44 & 1.00 \end{pmatrix}$$

39

Matrix algebra

- Important concepts/topics
  - Matrix addition/subtraction/multiplication
  - Special types of matrices
    - Correlation matrix
    - Sum of squares / Sum of cross products matrix
    - Variance / Covariance matrix

40

END SEGMENT

41

Lecture 11 ~ Segment 3

Estimation of coefficients

42

Estimation of coefficients

- The values of the coefficients (B) are estimated such that the model yields optimal predictions
  - Minimize the residuals!

43

Estimation of coefficients

- The sum of the squared (SS) residuals is minimized
  - SS.RESIDUAL = $\Sigma(\hat{Y} - Y)^2$

44

Estimation of coefficients

- Standardized and in matrix form, the regression equation is $\hat{Y} = B(X)$, where
  - $\hat{Y}$ is a [N x 1] vector
    - N = number of cases
  - B is a [k x 1] vector
    - k = number of predictors
  - X is a [N x k] matrix

45

Estimation of coefficients

- $\hat{Y} = B(X)$
  - To solve for B
  - Replace $\hat{Y}$ with Y
  - Conform for matrix multiplication:
    - Y = X(B)

46

Estimation of coefficients

- Y = X(B)
- Now let's make X square and symmetric
- To do this, pre-multiply both sides of the equation by the transpose of X, X^T

47

Estimation of coefficients

- Y = X(B) becomes
- $X^T(Y) = X^T(XB)$
- Now, to solve for B, eliminate X^TX
- To do this, pre-multiply by the inverse, $(X^TX)^{-1}$

48

Estimation of coefficients

- $X^TY = X^T(XB)$ becomes
- $(X^TX)^{-1}(X^TY) = (X^TX)^{-1}(X^TXB)$
  - Note that $(X^TX)^{-1}(X^TX) = I$
  - And $IB = B$
- Therefore, $(X^TX)^{-1}(X^TY) = B$

49

Estimation of coefficients

- $B = (X^TX)^{-1}(X^TY)$

50

Estimation of coefficients

- $B = (X^TX)^{-1}(X^TY)$
- Let's use this formula to calculate B's from the raw data matrix used in the previous segment

51

Raw data matrix

$$X_{np} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix}$$

52

Row vector of sums

$$T_{1p} = 1_{1n} * X_{np} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} * \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix} = \begin{pmatrix} 34 & 35 & 34 \end{pmatrix}$$

53

Row vector of means

$$M_{1p} = T_{1p} * N^{-1} = \begin{pmatrix} 34 & 35 & 34 \end{pmatrix} * 10^{-1} = \begin{pmatrix} 3.4 & 3.5 & 3.4 \end{pmatrix}$$

54

Matrix of means

$$M_{np} = 1_{n1} * M_{1p} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} * \begin{pmatrix} 3.4 & 3.5 & 3.4 \end{pmatrix} = \begin{pmatrix} 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \end{pmatrix}$$

55

Matrix of deviation scores

$$D_{np} = X_{np} - M_{np} = \begin{pmatrix} 3 & 2 & 3 \\ 3 & 2 & 3 \\ 2 & 4 & 4 \\ 4 & 3 & 4 \\ 4 & 4 & 3 \\ 5 & 4 & 3 \\ 2 & 5 & 4 \\ 3 & 3 & 2 \\ 5 & 3 & 4 \\ 3 & 5 & 4 \end{pmatrix} - \begin{pmatrix} 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \\ 3.4 & 3.5 & 3.4 \end{pmatrix} = \begin{pmatrix} -.4 & -1.5 & -.4 \\ -.4 & -1.5 & -.4 \\ -1.4 & .5 & .6 \\ .6 & -.5 & .6 \\ .6 & .5 & -.4 \\ 1.6 & .5 & -.4 \\ -1.4 & 1.5 & .6 \\ -.4 & -.5 & -1.4 \\ 1.6 & -.5 & .6 \\ -.4 & 1.5 & .6 \end{pmatrix}$$

56

SS & SP matrix

$S_{xx} = D_{pn}^T * D_{np} =$

$$\begin{pmatrix} -.4 & -.4 & -1.4 & .6 & .6 & 1.6 & -1.4 & -.4 & 1.6 & -.4 \\ -1.5 & -1.5 & .5 & -.5 & .5 & .5 & 1.5 & -.5 & -.5 & 1.5 \\ -.4 & -.4 & .6 & .6 & -.4 & -.4 & .6 & -1.4 & .6 & .6 \end{pmatrix} * \begin{pmatrix} -.4 & -1.5 & -.4 \\ -.4 & -1.5 & -.4 \\ -1.4 & .5 & .6 \\ .6 & -.5 & .6 \\ .6 & .5 & -.4 \\ 1.6 & .5 & -.4 \\ -1.4 & 1.5 & .6 \\ -.4 & -.5 & -1.4 \\ 1.6 & -.5 & .6 \\ -.4 & 1.5 & .6 \end{pmatrix}$$

$$= \begin{pmatrix} 10.4 & -2.0 & -.6 \\ -2.0 & 10.5 & 3.0 \\ -.6 & 3.0 & 4.4 \end{pmatrix}$$

SS & SP matrix

Since we used deviation scores:
Substitute S_{xx} for X^TX
Substitute S_{xy} for X^TY

Therefore,

$B = (S_{xx})^{-1}S_{xy}$

Estimation of coefficients

$B = (S_{xx})^{-1}S_{xy}$

$$B = \begin{pmatrix} 10.5 & 3.0 \\ 3.0 & 4.4 \end{pmatrix}^{-1} \begin{pmatrix} -2.0 \\ -.6 \end{pmatrix} = \begin{pmatrix} -0.19 \\ -0.01 \end{pmatrix}$$

Estimation of coefficients

```
Call:
lm(formula = demo$Y ~ demo$X1 + demo$X2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3011 -0.6855 -0.3091  0.6458  1.6909

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.086022   2.034669   2.008   0.0846 .
demo$X1     -0.188172   0.411447  -0.457   0.6613
demo$X2     -0.008065   0.635598  -0.013   0.9902
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.196 on 7 degrees of freedom
Multiple R-squared: 0.03665,  Adjusted R-squared: -0.2386
F-statistic: 0.1332 on 2 and 7 DF,  p-value: 0.8775
```

END SEGMENT

61

END LECTURE 11

62

Statistics One

Lecture 12
The General Linear Model (GLM)

1

Two segments

- The General Linear Model (GLM)
- Dummy coding

2

Lecture 12 ~ Segment 1

The General Linear Model (GLM)

3

General Linear Model (GLM)

- GLM is the mathematical framework used in many common statistical analyses, including multiple regression and ANOVA
  - ANOVA is typically presented as distinct from multiple regression but it IS a multiple regression

4

Characteristics of GLM

- *Linear*: pairs of variables are assumed to have linear relations
- *Additive*: if one set of variables predict another variable, the effects are thought to be additive

5

Characteristics of GLM

- BUT! This does not preclude testing non-linear or non-additive effects

6

Characteristics of GLM

- GLM can accommodate such tests, for example, by
  - Transformation of variables
    - Transform so non-linear becomes linear
  - Moderation analysis
    - Fake the GLM into testing non-additive effects

7

GLM example

- Simple regression
  - $Y = B_0 + B_1X_1 + e$
  - Y = faculty salary
  - X1 = years since PhD

8

GLM example

- Multiple regression
  - $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$
  - Y = faculty salary
  - X1 = years since PhD
  - X2 = number of publications
  - X3 = (years x pubs)

9

GLM example

- One-way ANOVA
  - $Y = B_0 + B_1X_1 + e$
  - Y = faculty salary
  - X1 = gender

10

GLM example

- Factorial ANOVA
  - $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$
  - Y = faculty salary
  - X1 = gender
  - X2 = race
  - X3 = interaction (gender x race)

11

Analysis of Variance (ANOVA)

- Appropriate when the predictors (IVs) are all categorical and the outcome (DV) is continuous
  - Most common application is to analyze data from randomized experiments

12

Analysis of Variance (ANOVA)

- More specifically, randomized experiments that generate more than 2 means
  - If only 2 means then use:
    - Independent t-test
    - Dependent t-test

13

General Linear Model (GLM)

- GLM is the mathematical framework used in many common statistical analyses, including multiple regression and ANOVA
  - ANOVA is typically presented as distinct from multiple regression but it IS a multiple regression

14

Characteristics of GLM

- *Linear*: pairs of variables are assumed to have linear relations
- *Additive*: if one set of variables predict another variable, the effects are thought to be additive

15

END SEGMENT

16

Lecture 12 ~ Segment 2

Dummy coding

17

Dummy coding

- A system to code categorical predictors in a regression analysis

Dummy coding

- Example
  - IV: Area of research in a Psychology department
    - Cognitive
    - Clinical
    - Developmental
    - Social
  - DV: Number of publications

Dataframe

ProfID	Group	Pubs
NU	Cognitive	83
ZH	Clinical	74
MK	Developmental	80
RH	Social	68

Dummy coding

	D1	D2	D3
Cognitive	0	0	0
Clinical	1	0	0
Developmental	0	1	0
Social	0	0	1

Dataframe

ProfID	Group	Pubs	D1	D2	D3
NU	Cognitive	83	0	0	0
ZH	Clinical	74	1	0	0
MK	Developmental	80	0	1	0
RH	Social	68	0	0	1

Summary statistics

Group	M	SD	N
Cognitive	93.31	29.48	13
Clinical	60.67	11.12	8
Developmental	103.50	23.64	6
Social	70.13	21.82	9
Total	81.69	27.88	36

23

Regression model

- $\hat{Y} = B_0 + B_1(D1) + B_2(D2) + B_3(D3)$

Coefficients

	B	SE	B	t	p
	93.31	6.50	0	14.37	<.001
D1 (Clinical)	-32.64	10.16	-.51	-3.21	.003
D2 (Devel)	10.19	11.56	.14	0.88	.384
D3 (Social)	-23.18	10.52	-.35	-2.20	.035

25

Unweighted effects coding

	C1	C2	C3
Cognitive	-1	-1	-1
Clinical	1	0	0
Developmental	0	1	0
Social	0	0	1

Coefficients

	B	SE	B	t	p
	81.90	4.06	0	14.37	<.001
D1 (Clinical)	-21.23	6.85	-.51	-3.21	.003
D2 (Devel)	21.60	7.88	.14	0.88	.384
D3 (Social)	-11.78	7.12	-.35	-2.20	.035

27

Weighted effects coding

	C1	C2	C3
Cognitive	$-N_{Clin}/N_{Cog}$	$-N_{Dev}/N_{Cog}$	$-N_{Soc}/N_{Cog}$
Clinical	N_{Clin}/N_{Cog}	0	0
Developmental	0	N_{Dev}/N_{Cog}	0
Social	0	0	N_{Soc}/N_{Cog}

Segment summary

- Dummy coding
  - A system to code categorical predictors in a regression analysis

END SEGMENT

30

END LECTURE 12

31

Statistics One

Lecture 13
Moderation

1

Three segments

- Moderation Example 1
- Centering predictors
- Moderation Example 2

2

Lecture 13 ~ Segment 1

Moderation Example 1

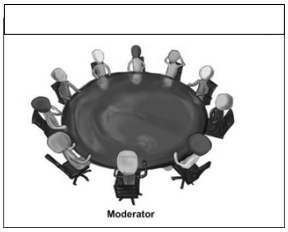
3

Moderation & Mediation

- Moderation and Mediation may sound alike but they are quite different
  - Moderation (Lecture 13)
  - Mediation (Lecture 14)
  - Both demonstrated in R (Lab 7)

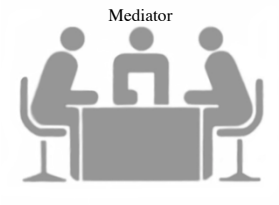
4

Moderation



5

Mediation



6

An example

- X: Experimental manipulation
  - Stereotype threat
- Y: Behavioral outcome
  - IQ test score
- Z: Moderator
  - Working memory capacity (WMC)

7

Moderation

- A moderator variable (Z) will enhance a regression model if the relationship between X and Y varies as a function of Z

8

Moderation

- Experimental research
  - The manipulation of an IV (X) causes change in a DV (Y)
  - A moderator variable (Z) implies that the effect of the IV on the DV (X on Y) is NOT consistent across the distribution of Z

9

Moderation

- Correlational research
  - Assume a correlaton between X and Y
  - A moderator variable (Z) implies that the correlation between X and Y is NOT consistent across the distribution of Z

10

Moderation

- If X and Y are correlated then we can use regression to predict Y from X
  - $Y = B_0 + B_1X + e$
  - CAUTION!
  - If there is a moderator, Z, then B_1 will NOT be representative across all Z
    - The relationship between X and Y is different at different levels of Z

11

Moderation model

- If both X and Z are continuous
  - $Y = B_0 + B_1X + B_2Z + B_3(X*Z) + e$

12

Moderation model

- If X is categorical* and Z is continuous
  - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z + B_4(D1*Z) + B_5(D2*Z) + e$

**3 levels of X*

13

How to test for moderation

- If both X and Z are continuous
  - Model 1: No moderation
    - $Y = B_0 + B_1X + B_2Z + e$
  - Model 2: Moderation
    - $Y = B_0 + B_1X + B_2Z + B_3(X*Z) + e$

14

How to test for moderation

- If X is categorical* and Z is continuous
  - Model 1: No moderation
    - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z + e$
  - Model 2: Moderation
    - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z + B_4(D1*Z) + B_5(D2*Z) + e$

15

How to test for moderation

- Compare Model 1 and Model 2 in terms of overall variance explained, that is, R^2
  - NHST available for this comparison
- Evaluate B values for predictors associated with the moderation effect
  - (X*Z)
  - (D1*Z) and (D2*Z)

16

Back to the example

- X: Experimental manipulation
  - Stereotype threat
- Y: Behavioral outcome
  - IQ test score
- Z: Moderator
  - Working memory capacity (WMC)

17

Simulated experiment & data

- Students completed a working memory task
- Students then randomly assigned to one of three experimental conditions
  - Explicit threat (n = 50)
  - Implicit threat (n = 50)
  - Control (n = 50)
- Students then completed an IQ test

18

Simulated experiment & data

- Experimental condition is categorical so dummy coding is required
  - Let the Control group be the referent
  - Let D1 = Explicit threat
  - Let D2 = Implicit threat

19

Results: Summary statistics

```
> describeBy(moderation.data, condition)
group: control
           var  n   mean    sd median trimmed   mad    min    max range skew kurtosis   se
subject      1 50  25.50 14.58  25.50   25.50 18.53   1.00  50.00 49.0 0.00    -1.27 2.06
condition*   2 50   1.00  0.00   1.00    1.00  0.00   1.00   1.00  0.0  NaN      NaN 0.00
IQ           3 50  97.88 20.93  99.50   97.47 25.20  46.00 141.00 95.0 0.04    -0.50 2.96
WM           4 50 102.18 18.79  99.50  100.55 20.02  71.00 159.00 88.0 0.73     0.41 2.66
WM_centered  5 50   3.18 18.79   0.42    1.47 20.02 -28.00  59.92 88.0 0.73     0.41 2.66
D1           6 50   0.00  0.00   0.00    0.00  0.00   0.00   0.00  0.0  NaN      NaN 0.00
D2           7 50   0.00  0.00   0.00    0.00  0.00   0.00   0.00  0.0  NaN      NaN 0.00
------------------------------------------------------------ 
group: threat1
           var  n   mean    sd median trimmed   mad    min    max range skew kurtosis   se
subject      1 50  75.50 14.58  75.50   75.50 18.53  51.00 100.00  49 0.00    -1.27 2.06
condition*   2 50   2.00  0.00   2.00    2.00  0.00   2.00   2.00   0  NaN      NaN 0.00
IQ           3 50  92.16 13.79  92.00   91.75 11.86  21.00  83.00  62 0.16    -0.25 1.95
WM           4 50 100.80 16.85  97.50   99.90 16.31  72.00 138.00  66 0.48    -0.56 2.38
WM_centered  5 50   1.72 16.85  -1.58    0.82 16.31 -27.08  38.92  66 0.48    -0.56 2.38
D1           6 50   1.00  0.00   1.00    1.00  0.00   1.00   1.00   0  NaN      NaN 0.00
D2           7 50   0.00  0.00   0.00    0.00  0.00   0.00   0.00   0  NaN      NaN 0.00
------------------------------------------------------------ 
group: threat2
           var  n   mean    sd median trimmed   mad    min    max range skew kurtosis   se
subject      1 50 125.50 14.58 125.50  125.50 18.53  101.00 150.00  49 0.00    -1.27 2.06
condition*   2 50   3.00  0.00   3.00    3.00  0.00   3.00   3.00   0  NaN      NaN 0.00
IQ           3 50  48.02 12.43  47.00   47.40 14.83  28.00  79.00  51 0.38    -0.56 1.76
WM           4 50  94.26 18.77  92.00   93.97 17.79  55.00 133.00  78 0.16    -0.71 2.65
WM_centered  5 50  -4.02 18.77  -7.00   -5.18 17.79 -44.08  33.92  78 0.16    -0.71 2.65
D1           6 50   0.00  0.00   0.00    0.00  0.00   0.00   0.00   0  NaN      NaN 0.00
D2           7 50   1.00  0.00   1.00    1.00  0.00   1.00   1.00   0  NaN      NaN 0.00
```

20

Results: Summary statistics

```
     correlations         group
[1,] "0.107982730191314" "IQ_WM_correlation_control"
[2,] "0.723109514848375" "IQ_WM_correlation_threat1"
[3,] "0.677291652202358" "IQ_WM_correlation_threat2"
```

21

Results: Model 1

```
> model1=lm(IQ ~ WM + D1 + D2)
> summary(model1)

Call:
lm(formula = IQ ~ WM + D1 + D2)

Residuals:
    Min      1Q  Median      3Q     Max
-47.339  -7.294   0.744   7.688  42.424

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.78635    7.14360   8.369 4.30e-14 ***
WM            0.37281    0.06688   5.575 1.16e-07 ***
D1          -45.28552    2.94638 -15.343  < 2e-16 ***
D2          -46.90735    2.99218 -15.677  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.72 on 146 degrees of freedom
Multiple R-squared:  0.7246, Adjusted R-squared:  0.719
F-statistic: 128.1 on 3 and 146 DF,  p-value: < 2.2e-16
```

22

Results: Model 2

```
> model2=lm(IQ ~ WM + D1 + D2 + (WM * D1) + (WM * D2))
> summary(model2)

Call:
lm(formula = IQ ~ WM + D1 + D2 + (WM * D1) + (WM * D2))

Residuals:
    Min      1Q  Median      3Q     Max
-50.414  -7.181   0.420   8.196  40.864

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.5851    11.3576   7.535 4.95e-12 ***
WM            0.1203     0.1094   1.100  0.27303
D1          -93.8952    16.8573  -5.571 1.52e-07 ***
D2          -79.8970    15.4772  -5.162 7.96e-07 ***
WM:D1         0.4716     0.1638   2.880  0.00459 **
WM:D2         0.3288     0.1547   2.125  0.03529 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.38 on 144 degrees of freedom
Multiple R-squared:  0.7409, Adjusted R-squared:  0.7319
F-statistic: 82.35 on 5 and 144 DF,  p-value: < 2.2e-16
```

23

Results: Model comparison

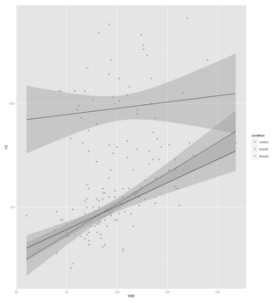
```
> anova(model1, model2)
Analysis of Variance Table

Model 1: IQ ~ WM + D1 + D2
Model 2: IQ ~ WM + D1 + D2 + (WM * D1) + (WM * D2)
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    146 31655
2    144 29784  2    1871.3 4.5238 0.01243 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

24

Results: Scatterplot

- Next slide depicts moderation visually

25



26

END SEGMENT

27

Lecture 13 ~ Segment 2

Centering predictors

28

Centering predictors

- To center means to put in deviation form
  - $X_C = X - M$
- Why center?
  - Two reasons
    - Conceptual
    - Statistical

Centering predictors

- Conceptual reason
  - Suppose
    - Y = child's verbal ability
    - X = mother's vocabulary
    - Z = child's age

Centering predictors

- Conceptual reason
  - The intercept, B_0, is the predicted score on Y when all predictors (X, Z) are zero
  - If X = zero or Z = zero is meaningless, or impossible, then B_0 will be difficult to interpret
  - In contrast, if X = zero and Z = zero, are the average then B_0 is easy to interpret

Centering predictors

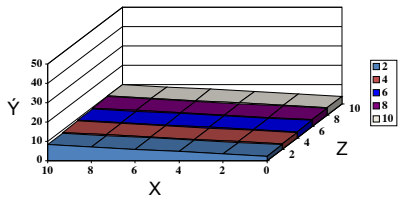
- Conceptual reason
  - The regression coefficient B_1 is the slope for X assuming an average score on Z
  - No moderation effect implies that B_1 is consistent across the entire distribution of Z

Centering predictors

- Conceptual reason
  - In contrast, a moderation effect implies that B_1 is NOT consistent across the entire distribution of Z
  - Where in the distribution of Z is B_1 most representative of the relationship between X & Y?
  - Let's look at this graphically…

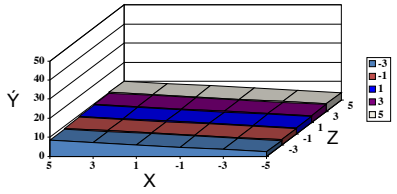
Uncentered, Additive

Ý = 2 +.6(X) + .2(Z)



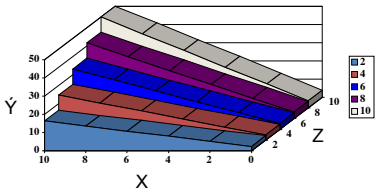
Centered, Additive

Ý = 6 +.6(X) + .2(Z)

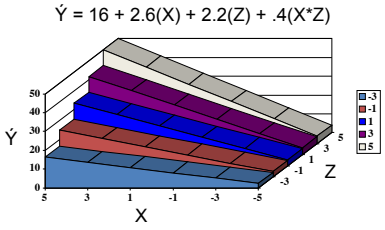


Uncentered, Moderation

Ý = 2 +.6(X) + .2(Z) + .4(X*Z)



Centered, Moderation



Centering predictors

- Statistical reason
  - The predictors, X and Z, can become highly correlated with the product, (X*Z)
    - *Multicolinearity*: when two predictor variables in a GLM are so highly correlated that they are essentially redundant and it becomes difficult to estimate B values associated with each predictor

Segment Summary

- Centering predictors
  - Convert raw scores to deviation scores
    - $X_C = X - M$
- Reasons for centering
  - Conceptual
    - Regression constant will be more meaningful
  - Statistical
    - Avoid multicolinearity

END SEGMENT

40

Lecture 13 ~ Segment 3

Moderation Example 2

41

Back to the example

- X: Experimental manipulation
  - Stereotype threat
- Y: Behavioral outcome
  - IQ test score
- Z: Moderator
  - Working memory capacity (WMC)

42

How to test for moderation

- If X is categorical* and Z is continuous
  - Model 1: No moderation
    - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z + e$
  - Model 2: Moderation
    - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z + B_4(D1*Z) + B_5(D2*Z) + e$

43

WAIT! Center continuous predictor

- If X is categorical* and Z is continuous
  - Model 1: No moderation
    - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z.center + e$
  - Model 2: Moderation
    - $Y = B_0 + B_1(D1) + B_2(D2) + B_3Z.center + B_4(D1*Z.center) + B_5(D2*Z.center) + e$

44

Simulated experiment & data

- Students completed a working memory task
- Students then randomly assigned to one of three experimental conditions
  - Explicit threat (n = 50)
  - Implicit threat (n = 50)
  - Control (n = 50)
- Students performed an IQ test

45

Simulated data

- Experimental condition is categorical so dummy coding is required
  - Let the Control group be the referent
  - Let D1 = Explicit threat
  - Let D2 = Implicit threat

46

Results: Model 1

```
> model1=lm(IQ ~ WM + D1 + D2)
> summary(model1)

Call:
lm(formula = IQ ~ WM + D1 + D2)

Residuals:
     Min       1Q   Median       3Q      Max
-47.339  -7.294   0.744   7.608  42.424

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59.78635    7.14360   8.369 4.30e-14 ***
WM            0.37281    0.06688   5.575 1.16e-07 ***
D1          -45.28552    2.94638 -15.343  < 2e-16 ***
D2          -46.90735    2.99218 -15.677  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.72 on 146 degrees of freedom
Multiple R-squared:  0.7246, Adjusted R-squared:  0.719
F-statistic: 128.1 on 3 and 146 DF,  p-value: < 2.2e-16
```

47

Results: Model 1, Centered

```
> model1.centered=lm(IQ ~ WM.centered + D1 + D2)
> summary(model1.centered)

Call:
lm(formula = IQ ~ WM.centered + D1 + D2)

Residuals:
     Min       1Q   Median       3Q      Max
-47.339  -7.294   0.744   7.608  42.424

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.72429    2.09267  46.220  < 2e-16 ***
WM.centered   0.37281    0.06688   5.575 1.16e-07 ***
D1          -45.28552    2.94638 -15.343  < 2e-16 ***
D2          -46.90735    2.99218 -15.677  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.72 on 146 degrees of freedom
Multiple R-squared:  0.7246, Adjusted R-squared:  0.719
F-statistic: 128.1 on 3 and 146 DF,  p-value: < 2.2e-16
```

48

Results: Model 2

```
> model2=lm(IQ ~ WM + D1 + D2 + (WM * D1) + (WM * D2))
> summary(model2)

Call:
lm(formula = IQ ~ WM + D1 + D2 + (WM * D1) + (WM * D2))

Residuals:
    Min      1Q  Median      3Q     Max
-50.414  -7.181   0.420   8.196  40.864

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.5851    11.3576   7.535 4.95e-12 ***
WM            0.1203     0.1094   1.100  0.27303
D1          -93.8952    16.8573  -5.523 1.52e-07 ***
D2          -79.8970    15.4772  -5.162 7.96e-07 ***
WM:D1         0.4716     0.1638   2.880  0.00459 **
WM:D2         0.3288     0.1547   2.125  0.03529 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.38 on 144 degrees of freedom
Multiple R-squared:  0.7409, Adjusted R-squared:  0.7319
F-statistic: 82.35 on 5 and 144 DF,  p-value: < 2.2e-16
```


Results: Model 2, Centered

```
> model2.centered=lm(IQ ~ WM.centered + D1 + D2 +
+                    (WM.centered * D1) + (WM.centered * D2))
> summary(model2.centered)

Call:
lm(formula = IQ ~ WM.centered + D1 + D2 + (WM.centered * D1) +
    (WM.centered * D2))

Residuals:
    Min      1Q  Median      3Q     Max
-50.414  -7.181   0.420   8.196  40.864

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     97.5078     2.8619  47.289  < 2e-16 ***
WM.centered      0.1203     0.1094   1.100  0.27303
D1             -46.3652     2.9038 -15.967  < 2e-16 ***
D2             -47.3223     2.9439 -16.075  < 2e-16 ***
WM.centered:D1   0.4716     0.1638   2.880  0.00459 **
WM.centered:D2   0.3288     0.1547   2.125  0.03529 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.38 on 144 degrees of freedom
Multiple R-squared:  0.7409, Adjusted R-squared:  0.7319
F-statistic: 82.35 on 5 and 144 DF,  p-value: < 2.2e-16
```


Results: Model comparison

```
> anova(model1, model2)
Analysis of Variance Table

Model 1: IQ ~ WM + D1 + D2
Model 2: IQ ~ WM + D1 + D2 + (WM * D1) + (WM * D2)
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    146 31655
2    144 29784  2    1871.3 4.5238 0.01243 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Results: Model comparison, Centered

```
> anova(model1.centered, model2.centered)
Analysis of Variance Table

Model 1: IQ ~ WM.centered + D1 + D2
Model 2: IQ ~ WM.centered + D1 + D2 + (WM.centered * D1) + (WM.centered *
    D2)
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    146 31655
2    144 29784  2    1871.3 4.5238 0.01243 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

END SEGMENT

53

END LECTURE 13

54

Statistics One

Lecture 14
Mediation

1

Two segments

- Standard approach
- Path analysis

2

Lecture 14 ~ Segment 1

Mediation: Standard approach

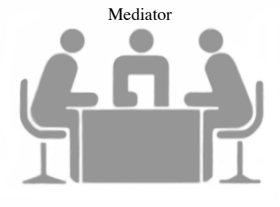
3

Mediation

- Mediation and moderation may sounds alike but they are quite different
  - Moderation (Lecture 13)
  - Mediation (Lecture 14)
  - Both demonstrated in R (Lab 7)

4

Mediation



5

An example

- X: Experimental manipulation
  - Stereotype threat
- Y: Behavioral outcome
  - IQ score
- M: Mediator (Mechanism)
  - Working memory capacity (WMC)

6

Mediation

- A mediation analysis is typically conducted to better understand an observed effect of an IV on a DV or a correlation between X and Y
  - Why, and how, does stereotype threat influence IQ test performance?

7

Mediation

- If X and Y are correlated then we can use regression to predict Y from X
  - $Y = B_0 + B_1X + e$

8

Mediation

- If X and Y are correlated BECAUSE of the mediator M, then (X → M → Y):
  - $Y = B_0 + B_1M + e$
    &
  - $M = B_0 + B_1X + e$

9

Mediation

- If X and Y are correlated BECAUSE of the mediator M, and:
  - $Y = B_0 + B_1M + B_2X + e$
  - What will happen to the predictive value of X
  - In other words, will B_2 be significant?

10

Mediation

- A mediator variable (M) accounts for some or all of the relationship between X and Y
  - *Some*: Partial mediation
  - *All*: Full mediation

11

Mediation

- CAUTION!
  - Correlation does not imply causation!
  - In other words, there is a BIG difference between statistical mediation and true causal mediation

12

How to test for mediation

- Run three regression models
  - lm(Y ~ X)
  - lm(M ~ X)
  - lm(Y ~ X + M)

13

How to test for mediation

- Run three regression models
  - lm(Y ~ X)
    - Regression coefficient for X should be significant
  - lm(M ~ X)
    - Regression coefficient for X should be significant

14

How to test for mediation

- Run three regression models
  - lm(Y ~ X + M)
    - Regression coefficient for M should be significant
    - Regression coefficient for X?

15

Back to the example

- X: Experimental manipulation
  - Stereotype threat
- Y: Behavioral outcome
  - IQ score
- M: Mediator (Mechanism)
  - Working memory capacity (WMC)

16

Simulated experiment & data

- Students randomly assigned to one of two experimental conditions
  - Threat
  - Control
- Students completed a working memory task
- Students completed an IQ test

17

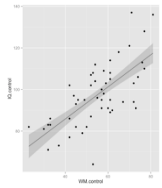
Results

```
> indirect
$`Mod1: Y~X`
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    97.32   2.070078 46.9991086 4.965625e-09
predthreat    -11.00   2.928388 -3.7563342 2.927777e-04

$`Mod2: Y~X+M`
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 55.997696 4.64483367 12.057987 5.303598e-21
predthreat  -2.407489 2.31641713 -1.039316 3.012416e-01
med          0.752489 0.07999527  9.405669 2.576515e-15

$`Mod3: M~X`
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    54.92   1.998708 28.494581 9.487423e-50
predthreat    -11.42   2.688807 -4.2485841 4.906301e-05

$Indirect.Effect
[1] -8.592511

$SE
[1] 2.219233

$z.value
[1] -3.871839

$N
[1] 100
```

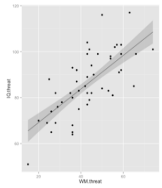
18

Results

Control group



Threat group



19

Interpretation

- Full mediation
  - The direct effect is no longer significant after adding the mediator into the regression equation
  - The Sobel test is significant

20

END SEGMENT

21

Lecture 14 ~ Segment 2

Mediation: Path analysis approach

22

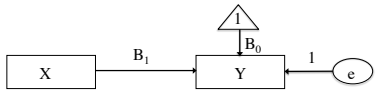
Mediation

- Mediation analyses are typically illustrated using "path models"
  - Rectangles: Observed variables (X, Y, M)
  - Circles: Unobserved variables (e)
  - Triangles: Constants
  - Arrows: Associations (more on these later)

23

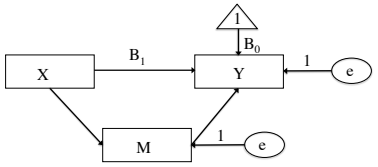
Path model

- $Y = B_0 + B_1X + e$



24

Path model with a mediator



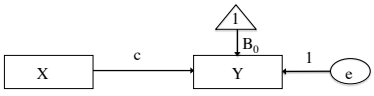
25

Path model with a mediator

- To avoid confusion, let's label the paths
  - a: Path from X to M
  - b: Path from M to Y
  - c: Direct path from X to Y (before including M)
  - c': Direct path from X to Y (after including M)
  - Note: (a*b) is known as the indirect path

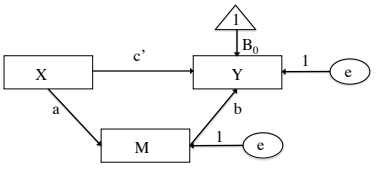
26

Path model



27

Path model with a mediator



28

How to test for mediation

- Three regression equations can now be re-written with new notation:
  - $Y = B_0 + c(X) + e$
  - $Y = B_0 + c'(X) + b(M) + e$
  - $M = B_0 + a(X) + e$

29

How to test for mediation

- The Sobel test

$$z = (B_a * B_b) / SQRT[(B_a^2 * SE_b^2) + (B_b^2 * SE_a^2)]$$

  - The null hypothesis
    - The indirect effect is zero
    - $(B_a * B_b) = 0$

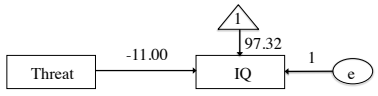
30

Results

```
> indirect
$`Mod1: Y~X`

             Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    97.32   2.079678 46.8991186 4.965623e-69
predthreat    -11.00   2.928380 -3.756342 2.927777e-04

$`Mod2: Y~X+M`

             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 55.997696 4.64403367 12.057987 5.303590e-21
predthreat  -2.487489 2.31641713 -1.073916 3.012416e-01
med          0.752489 0.07999527  9.405669 2.576515e-15

$`Mod3: M~X`

             Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    54.92   1.908708 28.8945981 9.487423e-50
predthreat    -11.42   2.688007 -4.248501 4.906391e-05

$indirect.effect
[1] -8.592511

$SE
[1] 2.219233

$z.value
[1] -3.871839

$N
[1] 100
```

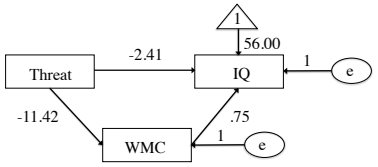
31

Path model



32

Path model with a mediator



33

Mediation: Final comments

- Here we used path analysis to *illustrate* the mediation analysis
- It is also possible to test for mediation using a statistical procedure called:
  - Structural Equation Modeling (SEM)

34

END SEGMENT

35

END LECTURE 14

36

Statistics One

Lecture 15
Student's t-test

1

Three segments

- Introduction
- Dependent t-tests
- Independent t-tests

2

Lecture 15 ~ Segment 1

Introduction

3

Introduction

- From multiple regression to t-tests?!
  - This is an unusual progression for an introduction to statistics
  - So why take this approach?

4

Introduction

- To reiterate the lesson from Lecture 1
  - Nothing beats a simple elegant randomized controlled experiment!

5

Introduction

- The examples discussed in multiple regression were complicated, considering the limitations placed on the final interpretations, for example,
  - The slope for X is B
  - But if you add another X then the slope changes!

6

Introduction

- The examples discussed in multiple regression were complicated, considering the limitations placed on the final interpretations, for example,
  - X and Y are correlated
  - But if you add a moderator variable
  - X and Y are not correlated!

7

Introduction

- Let's assume a simple experimental design
  - Independent variable
    - Vaccine
    - Placebo
  - Dependent variable
    - Rate of polio

8

Introduction

- Two means can be compared using a t-test
  - NHST can be conducted, yielding a p-value
  - Effect size can also be calculated
  - Confidence intervals around the sample means can also be reported

9

Introduction

- In this lecture, 4 tests, each compare means
  - z-test
  - t-test (single sample)
  - t-test (dependent)
  - t-test (independent)

10

Introduction

- Why is it called Student's t-test?

11

Introduction

- Developed by William Gossett in 1908
  - To monitor the quality of stout beer at the Guiness brewery in Dublin, Ireland
  - Management at Guiness considered their process a secret so they convinced Gossett to publish his work using the pen name "Student"

12

Introduction

- z = (Observed – Expected) / SE
- t = (Observed – Expected) / SE
  - SE: Standard error

13

When to use z and t?

- z-test
  - When comparing a sample mean to a population mean and the standard deviation of the population is known
- Single sample t-test
  - When comparing a sample mean to a population mean and the standard deviation of the population is not known

14

When to use z and t?

- Dependent t-test
  - When evaluating the difference between two related samples
- Independent t-test
  - When evaluating the difference between two independent samples

15

Observed, Expected, and SE

	Observed	Expected	SE
z	Sample mean	Population mean	SE of the mean
t (single sample)	Sample mean	Population mean	SE of the mean
t (dependent)	Sample mean of difference scores	Population mean of difference scores	SE of the mean difference
t (independent)	Difference between two sample means	Difference between two population means	SE of the difference between Ms

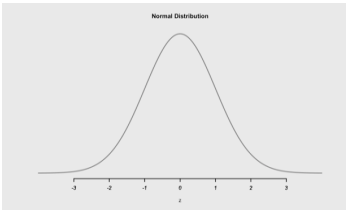
16

p-values for z and t

- Exact p-value depends on:
  - Directional or non-directional test
  - Degrees of freedom (df)
    - Different t-distributions for different sample sizes

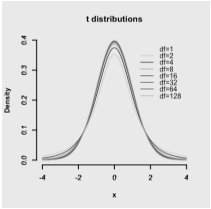
17

z distribution



18

Family of t distributions



19

Degrees of freedom (df)

	df
z	NA
t (single sample)	N-1
t (dependent)	N-1
t (independent)	(N1 – 1) + (N2 – 1)

20

Segment summary

- z = (Observed – Expected) / SE
- t = (Observed – Expected) / SE
  - SE: Standard error

21

Segment summary

- z-test
  - When comparing a sample mean to a population mean and the standard deviation of the population is known
- Single sample t-test
  - When comparing a sample mean to a population mean and the standard deviation of the population is not known

22

Segment summary

- Dependent t-test
  - When evaluating the difference between two related samples
- Independent t-test
  - When evaluating the difference between two independent samples

23

END SEGMENT

24

Lecture 15 ~ Segment 2

Dependent t-tests

25

Dependent t-test

- Also known as paired samples t-test
  - Appropriate when the same subjects are being compared
    - For example, pre/post design
  - Or when two samples are matched at the level of individual subjects
    - Allowing for a difference score to be calculated

26

Dependent t-test

- A thorough analysis will include
  - t-value
  - p-value
  - Cohen's d (effect size)
  - Confidence interval (interval estimate)

27

Dependent t-test

- t-value
  - t = (Observed – Expected) / SE
  - t = (M – 0) / SE = M / SE

28

Dependent t-test

- p-value
  - Based on t-value and the t-distribution
  - Directional or non-directional test

29

Dependent t-test

- Cohen's d
  - d = M / SD

30

Dependent t-test

- Confidence interval
  - Upper bound = M + t (SE)
  - Lower bound = M – t (SE)
  - t-value depends on level of confidence and t-distribution

31

Dependent t-test

- Examples
  - Wine ratings
  - Working memory training

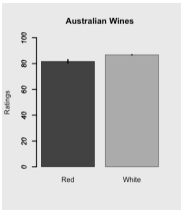
32

Dependent t-test

- Wine ratings
  - Each wine expert rated two wines, one red and one white
  - We can therefore compare the means
  - Australia was the only country that provided a normal distribution for both red and white

33

Dependent t-test



34

Dependent t-test

```
> # t-test
> t.test(australia.red, australia.white, paired=T)

 Paired t-test

data:  australia.red and australia.white
t = -8.0217, df = 99, p-value = 2.156e-12
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.300162 -3.801467
sample estimates:
mean of the differences
              -5.050815

> cohensD(australia.red, australia.white, method='paired')
[1] 0.8021719
```

35

Dependent t-test

- Working memory training
  - Let's compare intelligence scores before and after training (pre/post)

36

Dependent t-test

37

Dependent t-test

```
> t.test(pre, post, paired=T)

 Paired t-test

data:  pre and post
t = -3.5691, df = 19, p-value = 0.002047
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.5654256 -0.9295024
sample estimates:
mean of the differences
             -2.247464

> cohensD(pre, post, method='paired')
[1] 0.7980851
```

38

Segment summary

- Dependent t-test (paired samples t-test)
  - Appropriate when the same subjects are being compared
    - For example, pre/post design
  - Or when two samples are matched at the level of individual subjects
    - Allowing for a difference score to be calculated

39

Segment summary

- A thorough analysis will include
  - t-value
  - p-value
  - Cohen's d (effect size)
  - Confidence interval (interval estimate)

40

END SEGMENT

41

Lecture 15 ~ Segment 3

Independent t-tests

42

Independent t-test

- Compares two independent samples
  - For example, males and females, control and experimental, patients and healthy controls, etc.

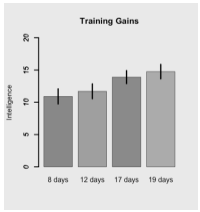
43

Independent t-test

- Example
  - Working memory training
    - Four independent groups trained for different amounts of time (8, 12, 17, or 19 days)

44

Working memory training



45

Independent t-test

- A thorough analysis will include
  - t-value
  - p-value
  - Cohen's d (effect size)
  - Confidence interval (interval estimate)

46

Independent t-test

- t-value
  - t = (Observed – Expected) / SE
  - t = (M1 – M2) / SE
  - SE = (SE1 + SE2) / 2

47

Independent t-test

- p-value
  - Based on t-value and the t-distribution
  - Directional or non-directional test

48

Independent t-test

- Cohen's d
  - d = (M1 – M2) / SD_{pooled}
  - SD_{pooled} = (SD1 + SD2) / 2

49

Independent t-test

- Confidence interval
  - Upper bound = M + t (SE)
  - Lower bound = M – t (SE)
  - t-value depends on level of confidence and t-distribution

50

Independent t-test

- Homogeneity of variance assumption
  - The pooled SD is appropriate only if the variances in the two groups are equivalent
  - If not then the homogeneity of variance assumption is violated
    - Simulations indicate this results in an increase in the probability of a Type I error

51

Independent t-test

- Homogeneity of variance assumption
  - How to detect a violation:
    - Conduct Levene's test
      - If significant then the homogeneity of variance assumption is violated

52

Independent t-test

- Homogeneity of variance assumption
  - What to do if violated?
    - Adjust df and p-value (Welch's procedure)
    - Use a non-parametric test (Lecture 24)

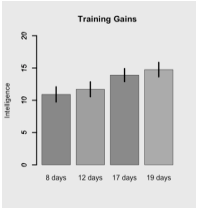
53

Back to the examples

- Example 1
  - Working memory training
    - Four independent groups trained for different amounts of time (8, 12, 17, or 19 days)

54

Working memory training



55

Results: Summary statistics

```
> describe(Days8)
  var  n  mean   sd median trimmed  mad  min   max range  skew kurtosis   se
1   1 20 10.91 2.62   11.3   10.98 2.64 5.36 15.65 10.29 -0.24    -0.69 0.59
> describe(Days12)
  var  n mean   sd median trimmed  mad  min   max range skew kurtosis   se
1   1 20 11.7 2.58  11.62   11.69 2.84 6.91 16.12   9.2 0.06    -1.06 0.58
> describe(Days17)
  var  n mean   sd median trimmed  mad  min   max range skew kurtosis  se
1   1 20 13.9 2.26   13.6    13.9 2.03 9.81 18.12  8.31  0.1    -0.84 0.5
> describe(Days19)
  var  n  mean  sd median trimmed mad   min   max range  skew kurtosis   se
1   1 20 14.75 2.5  15.31   14.71 2.2 10.37 19.24  8.87 -0.08    -0.98 0.56
```

56

Results: Levene's test

```
> leveneTest(WMT$IQ ~ WMT$condition)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.1405 0.9355
      76
```

57

Results: 8 vs. 12

```
> t.test(Days8, Days12, var.equal=T)

 Two Sample t-test

data:  Days8 and Days12
t = -0.9584, df = 38, p-value = 0.3439
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.456381  0.877821
sample estimates:
mean of x mean of y
 10.91359  11.70287

> cohensD(Days8, Days12, method='pooled')
[1] 0.3030849
```

58

Results: 8 vs. 17

```
> t.test(Days8, Days17, var.equal=T)

 Two Sample t-test

data:  Days8 and Days17
t = -3.8624, df = 38, p-value = 0.0004239
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.556556 -1.422647
sample estimates:
mean of x mean of y
 10.91359  13.90319

> cohensD(Days8, Days17, method='pooled')
[1] 1.221383
```

59

Results: 8 vs. 19

```
> t.test(Days8, Days19, var.equal=T)

 Two Sample t-test

data:  Days8 and Days19
t = -4.7351, df = 38, p-value = 3.026e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.482354 -2.198557
sample estimates:
mean of x mean of y
 10.91359  14.75404

> cohensD(Days8, Days19, method='pooled')
[1] 1.497378
```

60

Results: 12 vs. 17

```
> t.test(Days12, Days17, var.equal=T)

 Two Sample t-test

data:  Days12 and Days17
t = -2.868, df = 38, p-value = 0.006706
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.7534458 -0.6471965
sample estimates:
mean of x mean of y
 11.70287  13.90319

> cohensD(Days12, Days17, method='pooled')
[1] 0.9069322
```

61

Results: 12 vs. 19

```
> t.test(Days12, Days19, var.equal=T)

 Two Sample t-test

data:  Days12 and Days19
t = -3.7925, df = 38, p-value = 0.00052
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.679880 -1.422471
sample estimates:
mean of x mean of y
 11.70287  14.75404

> cohensD(Days12, Days19, method='pooled')
[1] 1.199278
```

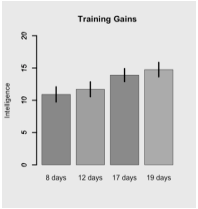
62

Results: 17 vs. 19

```
> t.test(Days17, Days19, var.equal=T)

 Two Sample t-test

data:  Days17 and Days19
t = -1.1287, df = 38, p-value = 0.2661
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.3768951  0.6751862
sample estimates:
mean of x mean of y
 13.90319  14.75404

> cohensD(Days17, Days19, method='pooled')
[1] 0.356931
```

63

Working memory training



64

Problems?

- Conducting multiple t-tests like that…
  - Is tedious
  - Increases the probability of Type I error
  - When there are more than two group means to compare, conduct Analysis of Variance (ANOVA)

65

END SEGMENT

66

END LECTURE 15

67

Statistics One

Lecture 16
Analysis of Variance (ANOVA)

1

Two segments

- One-way ANOVA
- Post-hoc tests

2

Lecture 16 ~ Segment 1

One-way ANOVA

3

Analysis of Variance (ANOVA)

- Appropriate when the predictors (IVs) are all categorical and the outcome (DV) is continuous
  - Most common application is to analyze data from randomized controlled experiments

4

Analysis of Variance (ANOVA)

- More specifically, randomized controlled experiments that generate more than two group means
  - If only two group means then use:
    - Independent t-test
    - Dependent t-test

5

Analysis of Variance (ANOVA)

- If more than two group means then use:
  - Between groups ANOVA
  - Repeated measures ANOVA

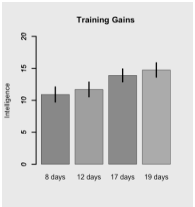
6

Example

- Working memory training
  - Four independent groups (8, 12, 17, 19)
    - IV: Number of training sessions
    - DV: IQ gain
    - Null hypothesis: All groups are equal

7

Working memory training



8

Analysis of Variance (ANOVA)

- ANOVA typically involves NHST
- The test statistic is the F-test (F-ratio)
  - F = (Variance between groups) / (Variance within groups)

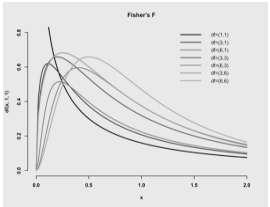
9

Analysis of Variance (ANOVA)

- Like the t-test and family of t-distributions
- The F-test has a family of F-distributions
  - The distribution to assume depends on
    - Number of subjects per group
    - Number of groups

10

Analysis of Variance (ANOVA)



11

One-way ANOVA

- F-ratio

F = between-groups variance / within-groups variance

$F = MS_{Between} / MS_{Within}$

$F = MS_{A} / MS_{S/A}$

12

One-way ANOVA

- $F = MS_A / MS_{S/A}$
- $MS_A = SS_A / df_A$
- $MS_{S/A} = SS_{S/A} / df_{S/A}$

13

One-way ANOVA

- $SS_A = n \Sigma(Y_j - Y_T)^2$
  - Y_j are the group means
  - Y_T is the grand mean

14

One-way ANOVA

- $SS_{S/A} = \Sigma(Y_{ij} - Y_j)^2$
  - Y_{ij} are individual scores
  - Y_j are the group means

15

One-way ANOVA

- $df_A = a - 1$
- $df_{S/A} = a(n - 1)$
- $df_{TOTAL} = N - 1$

16

Summary Table

Source	SS	df	MS	F
A	$n\,\Sigma(Y_j - Y_T)^2$	a - 1	SS_A/df_A	$MS_A/MS_{S/A}$
S/A	$\Sigma(Y_{ij} - Y_j)^2$	a(n -1)	$SS_{S/A}/df_{S/A}$	-----
Total	$\Sigma(Y_{ij} - Y_T)^2$	N - 1	-----	-----

17

Effect size

- $R^2 = \eta^2$ (eta-sqaured)
- $\eta^2 = SS_A / SS_{Total}$

18

Assumptions

- DV is continuous (interval or ratio variable)
- DV is normally distributed
- Homogeneity of variance
  - Within-groups variance is equivalent for all groups
    - Levene's test

19

Homogeneity of variance

- If Levene's test is significant then homogeneity of variance assumption has been violated
  - Conduct pairwise comparisons using a restricted error term

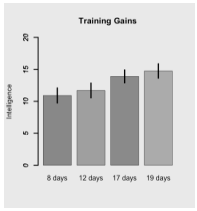
20

Example

- Working memory training
  - Four independent groups (8, 12, 17, 19)
    - IV: Number of training sessions
    - DV: IQ gain
    - Null hypothesis: All groups are equal

21

Working memory training



22

Working memory training

```
> anova.WMT <- aov(WMT$IQ ~ WMT$condition)
> summary(anova.WMT)
              Df Sum Sq Mean Sq F value   Pr(>F)
WMT$condition  3  196.1   65.36   10.49 7.47e-06 ***
Residuals     76  473.4    6.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Post hoc (Tukey)
> TukeyHSD(anova.WMT)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = WMT$IQ ~ WMT$condition)
```

23

Working memory training

```
> anova.WMT <- aov(WMT$IQ ~ WMT$condition)
> summary(anova.WMT)
              Df Sum Sq Mean Sq F value   Pr(>F)
WMT$condition  3  196.1   65.36   10.49 7.47e-06 ***
Residuals     76  473.4    6.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Post hoc (Tukey)
> TukeyHSD(anova.WMT)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = WMT$IQ ~ WMT$condition)

$`WMT$condition`
                    diff       lwr       upr     p adj
17 days-12 days  2.205  0.131797  4.278203 0.0326836
19 days-12 days  3.050  0.976797  5.123203 0.0013069
8 days-12 days  -0.790 -2.863203  1.283203 0.7492938
19 days-17 days  0.845 -1.228203  2.918203 0.7083557
8 days-17 days  -2.995 -5.068203 -0.921797 0.0016487
8 days-19 days  -3.840 -5.913203 -1.766797 0.0000354
```

24

Results from t-test: 12 vs. 17

```
> t.test(Days12, Days17, var.equal=T)

 Two Sample t-test

data:  Days12 and Days17
t = -2.868, df = 38, p-value = 0.006706
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.7534458 -0.6471965
sample estimates:
mean of x mean of y
 11.70287  13.90319

> cohensD(Days12, Days17, method='pooled')
[1] 0.9069322
```

25

Segment summary

- ANOVA is used to compare means, typically in experimental research
  - Categorical IV
  - Continuous DV

26

Segment summary

- ANOVA assumes homogeneity of variance
  - Evaluate with Levene's test

27

Segment summary

- Post-hoc tests, such as Tukey's procedure, allow for multiple pairwise comparisons without an increase in the probability of a Type I error

28

END SEGMENT

29

Lecture 16 ~ Segment 2

Post-hoc tests

30

Post-hoc tests

- Post-hoc tests, such as Tukey's procedure, allow for multiple pairwise comparisons without an increase in the probability of a Type I error

31

Post-hoc tests

- Many procedures are available; the degree to which p-values are adjusted varies according to procedure
  - Most liberal: No adjustment
  - Most conservative: Bonferroni procedure

32

NHST review

Experimenter Decision

Truth		Retain H_0	Reject H_0
	H_0 true	Correct Decision	Type I error (False alarm)
	H_0 false	Type II error (Miss)	Correct Decision

33

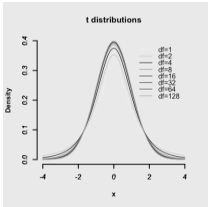
NHST review

Experimenter Decision

Truth		Retain H_0	Reject H_0
	H_0 true	Correct Decision	Type I error p = .05
	H_0 false	Type II error (Miss)	Correct Decision

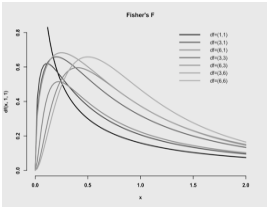
34

NHST review



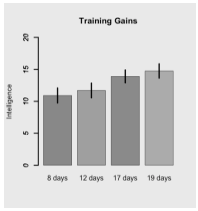
35

NHST review



36

Working memory training



37

Tukey's procedure

```
> anova.WMT <- aov(WMT$IQ ~ WMT$condition)
> summary(anova.WMT)
                Df Sum Sq Mean Sq F value   Pr(>F)
WMT$condition  3  196.1   65.36   10.49 7.47e-06 ***
Residuals     76  473.4    6.23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # Post hoc (Tukey)
> TukeyHSD(anova.WMT)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = WMT$IQ ~ WMT$condition)

$`WMT$condition`
                      diff       lwr       upr     p adj
17 days-12 days  2.205  0.131797  4.278203 0.0326836
19 days-12 days  3.050  0.976797  5.123203 0.0013069
8 days-12 days  -0.790 -2.863203  1.283203 0.7492938
19 days-17 days  0.845 -1.228203  2.918203 0.7083557
8 days-17 days  -2.995 -5.068203 -0.921797 0.0016487
8 days-19 days  -3.840 -5.913203 -1.766797 0.0000354
```

38

Results from t-test: 12 vs. 17

```
> t.test(Days12, Days17, var.equal=T)

 Two Sample t-test

data:  Days12 and Days17
t = -2.868, df = 38, p-value = 0.006706
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.7534458 -0.6471965
sample estimates:
mean of x mean of y
 11.70287  13.90319

> cohensD(Days12, Days17, method='pooled')
[1] 0.9069322
```

39

Bonferroni procedure

```
> p.adjust(.006706, method="bonferroni", 6)
[1] 0.040236
```

40

Comparison of procedures

Procedure	p-value for 12 vs. 17
Independent t-test	0.0067
Tukey	0.0327
Bonferroni	0.0402

41

Post-hoc tests

- Post-hoc tests, such as Tukey's procedure, allow for multiple pairwise comparisons without an increase in the probability of a Type I error
- Procedures vary from liberal to conservative

42

END SEGMENT

43

END LECTURE 16

44

Statistics One

Lecture 17
Factorial ANOVA

1

Two segments

- Factorial ANOVA
- Example

2

Lecture 17 ~ Segment 1

Factorial ANOVA

3

Factorial ANOVA

- Two Independent Variables (IVs)
- One Dependent Variable (DV)

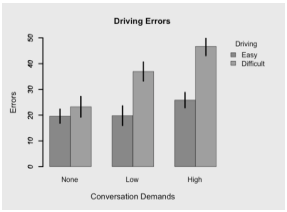
4

Example

- Suppose an experiment is conducted to examine the effect of talking on a mobile phone while driving
  - IV1: Driving difficulty
  - IV2: Conversation demand
  - DV: Driving errors

5

Example



6

Factorial ANOVA

- Three hypotheses can be tested:
  - More errors in the difficult simulator?
  - More errors with more demanding conversation?
  - More errors due to the interaction of driving difficulty and conversation demand?

7

Factorial ANOVA

- Three F ratios
  - F_A
  - F_B
  - F_{AxB}

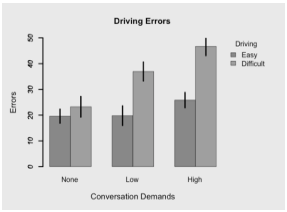
8

Factorial ANOVA

- *Main effect*: the effect of one IV averaged across the levels of the other IV

9

Example



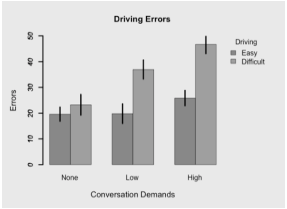
10

Factorial ANOVA

- *Interaction effect*: the effect of one IV depends on the other IV (the simple effects of one IV change across the levels of the other IV)

11

Example



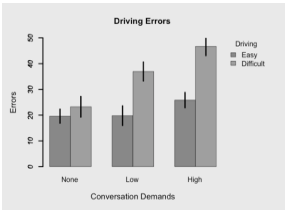
12

Factorial ANOVA

- *Simple effect*: the effect of one IV at a particular level of the other IV

13

Example



14

Factorial ANOVA

- Main effects and interaction effect are independent from one another

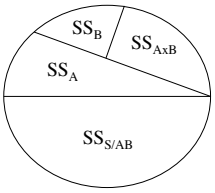
15

Factorial ANOVA

- Remember, factorial ANOVA is just a special case of multiple regression
  - It is a multiple regression with perfectly independent predictors (IVs)

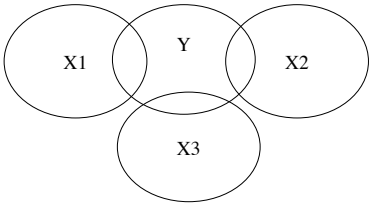
16

Partition SS in the DV



17

Independent predictor variables



18

Remember, GLM

- General Linear Model (GLM)
  - $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$
  - Y = DV
  - X1 = A
  - X2 = B
  - X3 = (A*B)

19

F ratios

- $F_A = MS_A / MS_{S/AB}$
- $F_B = MS_B / MS_{S/AB}$
- $F_{AxB} = MS_{AxB} / MS_{S/AB}$

20

MS

- $MS_A = SS_A / df_A$
- $MS_B = SS_B / df_B$
- $MS_{AxB} = SS_{AxB} / df_{AxB}$
- $MS_{S/AB} = SS_{S/AB} / df_{S/AB}$

21

df

- $df_A = a - 1$
- $df_B = b - 1$
- $df_{AxB} = (a -1)(b - 1)$
- $df_{S/AB} = ab(n - 1)$
- $df_{Total} = abn - 1 = N - 1$

22

Follow-up tests

- Main effects
  - Post-hoc tests
- Interaction
  - Analysis of simple effects
    - Conduct a series of one-way ANOVAs (or t-tests)

23

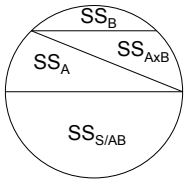
Effect size

- Complete η^2
  - $\eta^2 = SS_{effect} / SS_{total}$
- Partial η^2
  - $\eta^2 = SS_{effect} / (SS_{effect} + SS_{S/AB})$

24

Effect size (complete)

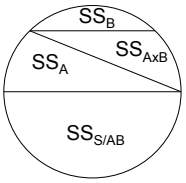
η^2 for the interaction = SS_{AxB} / SS_{Total}



25

Effect size (partial)

η^2 for the interaction = $SS_{AxB} / (SS_{AxB} + SS_{S/AB})$



26

Assumptions

- Assumptions underlying factorial ANOVA
  - DV is continuous (interval or ratio variable)
  - DV is normally distributed
  - Homogeneity of variance

27

Segment summary

- Factorial ANOVA
  - Three F-tests (F_A, F_B, F_{AxB})
  - Main effects
  - Interaction effect
  - Simple effects

28

Segment summary

- Factorial ANOVA
  - Effect size (complete and partial eta-squared)
  - Post-hoc tests (follow main effects)
  - Simple effects analyses (follow interaction)
  - Homogeneity of variance assumption
    - Levene's test

29

END SEGMENT

30

Lecture 17 ~ Segment 2

Factorial ANOVA
Example

31

Example

- Strayer and Johnson (2001) conducted an experiment to examine the effect of talking on a mobile phone while driving
- They tested subjects in a driving simulator

32

Example

- To manipulate driving difficulty, they simply made the driving course in the simulator more or less difficult

33

Example

- To manipulate conversation demand, they included two “talking” conditions:
  - In one condition the subject simply had to repeat what they heard on the other line of the phone

34

Example

- To manipulate conversation demand, they included two “talking” conditions:
  - In the other condition the subject had to think of and then say a word beginning with the last letter of the last word spoken on the phone
  - For example, if you hear “ship”, say a word that begins with the letter “p”, such as “peach”

35

Example

- IV1 = driving difficulty (easy, difficult)
- IV2 = conversation demand (none, low, high)
- DV = errors in driving simulator

36

Example

37

Results: Levene's test

```
> leveneTest(df$errors ~ df$driving * df$conversation)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   5  0.5206 0.7602
      114
```

38

Results: Factorial ANOVA

```
> summary(anova <- aov(df$errors ~ df$driving * df$conversation))
                          Df Sum Sq Mean Sq F value   Pr(>F)
df$driving                 1   5782    5782   94.64  < 2e-16 ***
df$conversation            2   4416    2208   36.14 6.98e-13 ***
df$driving:df$conversation 2   1639     820   13.41 5.86e-06 ***
Residuals                114   6965      61
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

39

Results: Simple effects

- Simple effect of A at each level of B
  - Effect of driving difficulty at each level of conversation demand
- Simple effect of B at each level of A
  - Effect of conversation demand at each level of driving difficulty

40

Example

41

Results: Simple effects

```
> t.test(none.easy, none.diff, var.equal=T)

 Two Sample t-test

data:  none.easy and none.diff
t = -1.5052, df = 38, p-value = 0.1405
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.55906  1.25906
sample estimates:
mean of x mean of y
    19.60     23.25

> cohensD(none.easy, none.diff)
[1] 0.475981
```

42

Results: Simple effects

```
> t.test(low.easy, low.diff, var.equal=T)

 Two Sample t-test

data:  low.easy and low.diff
t = -6.4625, df = 38, p-value = 1.324e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -22.52228 -11.77772
sample estimates:
mean of x mean of y
    19.80     36.95

> cohensD(low.easy, low.diff)
[1] 2.043623
```

43

Results: Simple effects

```
> t.test(high.easy, high.diff, var.equal=T)

 Two Sample t-test

data:  high.easy and high.diff
t = -8.9664, df = 38, p-value = 6.467e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.55742 -16.14258
sample estimates:
mean of x mean of y
    25.85     46.70

> cohensD(high.easy, high.diff)
[1] 2.835426
```

44

Segment summary

45

END SEGMENT

46

END LECTURE 17

47

Statistics One

Lecture 18
Repeated measures ANOVA

1

Two segments

- Repeated measures: Pros & Cons
- Repeated measures: Example

2

Lecture 18 ~ Segment 1

Repeated measures
Pros & Cons

3

Repeated measures: Pros & cons

- Pros
  - Less cost (fewer subjects required)
  - More statistical power
    - This is the important new concept

Repeated measures: Pros & cons

- Working memory training example
- Four independent groups (8, 12, 17, 19)
  - There were 20 subjects per group
  - Total N = 80

5

Repeated measures: Pros & cons

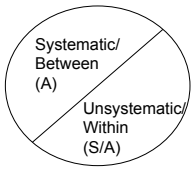
- Working memory training example
- Repeated measures design
  - N = 20

6

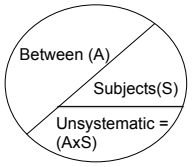
Repeated measures: Pros & cons

- More statistical power
  - Variance across subjects may be systematic
  - If so, it will not contribute to the error term

Between groups design (SS)



Repeated measures design (SS)



Error in a repeated measures design is the inconsistency of subjects from one condition to another

Therefore:

$F_A = MS_A / MS_{AxS}$

MS and F

- $MS_A = SS_A / df_A$
- $MS_{AxS} = SS_{AxS} / df_{AxS}$
- $F = MS_A / MS_{AxS}$

Repeated measures: Pros & cons

- Cons
  - Order effects
  - Counterbalancing
  - Missing data
  - Extra assumption

Counterbalancing

- Consider a simple design with just two conditions, A1 and A2
- One approach is a Blocked Design
  - Subjects are randomly assigned to one of two "order" conditions
    - A1, A2
    - A2, A1

Counterbalancing

- Another approach is a Randomized Design
  - Conditions are presented randomly in a mixed fashion
    - A2, A1, A1, A2, A2, A1, A2…..

Counterbalancing

- Now suppose a = 3 and a blocked design
- There are 6 possible orders (3!)
  - A1, A2, A3
  - A1, A3, A2
  - A2, A1, A3
  - A2, A3, A1
  - A3, A1, A2
  - A3, A2, A1

Counterbalancing

- To completely counterbalance, subjects would be randomly assigned to one of 6 order conditions
- The number of conditions needed to completely counterbalance becomes large with more conditions
  - 4! = 24
  - 5! = 120

Counterbalancing

- With many levels of the IV a better approach is to use a “Latin Squares” design
- Latin Squares designs aren’t completely counterbalanced but every condition appears at every position at least once

Counterbalancing

- For example, if a = 3, then
  - A1, A2, A3
  - A2, A3, A1
  - A3, A1, A2

Missing data

- Two issues to consider
  - *Relative amount* of missing data
  - *Pattern* of missing data

Missing data ~ Relative amount

- How much is a lot?
  - No hard and fast rules
  - A rule of thumb is
    - Less than 10% on any one variable, OK
    - Greater than 10%, not OK

Missing data ~ Pattern?

- Is the pattern random or lawful?
  - This can easily be detected
  - For any variable of interest (X) create a new variable (XM)
    - XM = 0 if X is missing
    - XM = 1 if X is not missing
  - Conduct a t-test with XM as the IV
  - If significant then pattern of missing data *may be* lawful

Missing data ~ Remedies

- Drop all cases without a perfect profile
  - Drastic
  - Use only if you can afford it
- Keep all cases and estimate the values of the missing data points
  - There are several options for how to estimate values

Sphericity assumption

- Homogeneity of variance
- Homogeneity of covariance

Sphericity assumption

- How to test?
  - Mauchly's test
  - If significant then report an adjusted p-value
    - Greenhouse-Geisser
    - Huyn-Feldt

Segment summary

- Pros
  - Less cost (fewer subjects required)
  - More statistical power
    - This is the important new concept

24

Segment summary

- Cons
  - Order effects
  - Counterbalancing
  - Missing data
  - Extra assumption

25

END SEGMENT

26

Lecture 18 ~ Segment 2

Repeated measures ANOVA
Example

27

Repeated measures: Pros & cons

- Working memory training example
- Four independent groups (8, 12, 17, 19)
  - There were 20 subjects per group
  - Total N = 80

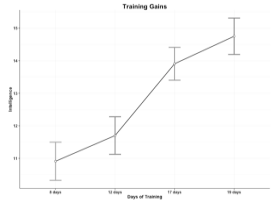
28

Repeated measures: Pros & cons

- Working memory training example
- Repeated measures design
  - N = 20

29

Working memory training



30

Common dataframe

subject	A1 (8)	A2 (12)	A3 (17)	A4 (19)
1				
2				
3				
4				
5				
6				
...				

31

R dataframe

subject	condition	IQ
1	A1 (8)	
1	A2 (12)	
1	A3 (17)	
1	A4 (19)	
2	A1 (8)	
2	A2 (12)	
...		

32

Results: ANOVA

```
> summary(anova <- aov(WM$IQ ~ WM$condition +
Error(factor(WM$subject)/WM$condition)))

Error: factor(WM$subject)
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 19  175.6   9.242

Error: factor(WM$subject):WM$condition
              Df Sum Sq Mean Sq F value   Pr(>F)
WM$condition  3  196.1   65.36   12.51 2.16e-06 ***
Residuals    57  297.8    5.22
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

33

Results: Post-hoc tests (Holm)

```
> with(WM, pairwise.t.test(IQ, condition, paired=T)) #all comp.

 Pairwise comparisons using paired t tests

data:  IQ and condition

        12 days 17 days 19 days
17 days 0.01924 -       -
19 days 0.00269 0.39572 -
8 days  0.39572 0.00237 0.00055

P value adjustment method: holm
```

34

Results: Post-hoc tests (Bonferroni)

```
> with(WM, pairwise.t.test(IQ, condition, paired=T, p.adjust.method="bonferroni")) #all comp.

 Pairwise comparisons using paired t tests

data:  IQ and condition

        12 days 17 days 19 days
17 days 0.03910 -       -
19 days 0.00405 1.00000 -
8 days  1.00000 0.00293 0.00054

P value adjustment method: bonferroni
```

35

Results: Paired t-test 12 vs. 17

```
> t.test(Days12, Days17, paired=T)

 Paired t-test

data:  Days12 and Days17
t = -3.0549, df = 19, p-value = 0.006517
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.7157116 -0.6942884
sample estimates:
mean of the differences
                 -2.205

> cohensD(Days12, Days17)
[1] 0.9087788
```

36

Comparison of procedures

Procedure	p-value for 12 vs. 17
Paired t-test	0.0065
Holm	0.0192
Bonferroni	0.0391

37

Repeated measures ANOVA

- Appropriate when comparing group means
  - Three or more group means
  - Same subjects tested in each condition
  - F-test
  - Post-hoc testss

38

END SEGMENT

39

END LECTURE 18

40

Statistics One

Lecture 21
Assumptions Revisited

1

Two segments

- Assumptions
- Transformations

2

Lecture 21 ~ Segment 1

Assumptions

3

Review of main assumptions

- Correlation
  - Normal distribution in X and Y
  - Linear relationship between X and Y
  - Homoscedasticity

  - Plot histograms and scatterplots
  - Print summary statistics

Review of main assumptions

- Regression
  - Normal distribution in Y
  - Linear relationship between X and Y
  - Homoscedasticity
  - Correlations among predictor variables all < .80
    - No multicolinearity
  - Plot histograms and scatterplots and residuals
  - Print summary statistics

Review of main assumptions

- t-tests
  - Normal distribution in Y (DV)
  - Homogeneity of variance
  - Equivalent sample size
  - Levene's test

Review of main assumptions

- Between groups and Factorial ANOVA
  - Normal distribution in Y (DV)
  - Homogeneity of variance
  - Equivalent sample size
  - Levene's test

Review of main assumptions

- Repeated measures ANOVA
  - Normal distribution in Y (DV)
  - Sphericity assumption
    - Homogeneity of variance
    - Homogeneity of covariance
  - Equivalent sample size
  - Levene's test
  - Mauchly's test

Review of main assumptions

- Chi-square
  - Independence
  - Adequate expected cell counts

Review of main assumptions

- Two primary constraints of the assumptions
  - Normal distribution in Y (DV)
  - Linear relationship between predictor variables and outcome variable (see Lecture 23)

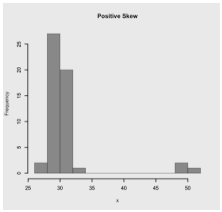
Review of main assumptions

- Normal distribution in Y (DV)
  - How to test
    - Histograms and summary statistics
    - Q-Q plots
    - Empirical tests, such as D'Agostino's K^2 test

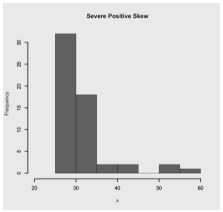
Review of main assumptions

- Normal distribution in Y (DV)
  - How to test
    - Histograms and summary statistics
    - Look for extreme skew and/or kurtosis and/or outliers (for example, cases +/- 3 SDs from the mean)
    - Rule of thumb is (skew > 3) and/or (kurtosis > 10) indicates a non-normal distribution

Positive skew with outliers



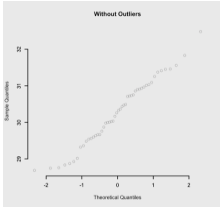
Positive skew



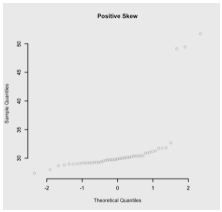
Review of main assumptions

- Normal distribution in Y (DV)
  - How to test
    - Q-Q plot
    - A plot of the sorted values from the data set against the expected values of the corresponding quantiles from the standard normal distribution.
    - If the distribution is normal then the plotted points should approximately lie on a straight line.

Normal distribution



Positive skew



Segment summary

- The inferential statistics covered in this course involve several assumptions
- Two primary assumptions
  - Normal distribution in the outcome Y (DV)
  - Linear relationship between predictors and Y

Segment summary

- Normal distribution in Y (DV)
  - How to test
    - Histograms and summary statistics
    - Q-Q plots
    - Empirical tests, such as D'Agostino's K^2 test

END SEGMENT

20

Lecture 21 ~ Segment 2

Transformations

21

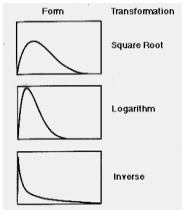
Transformations

- If a distribution is not normal then it is sometimes possible to transform the data in an attempt to make the distribution normal
- A transformation is a single function applied to all data points in the distribution (for example, square root)
- The rank order of cases should remain the same but the distance between cases may change

Transformations

- Most common transformations for positive skew
  - Square root
  - Logarithm
  - Inverse

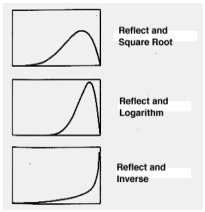
Transformations



Transformations

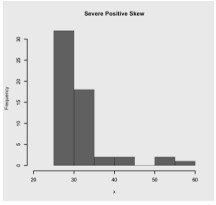
- Most common transformations for negative skew
  - Reflect and Square root
  - Reflect and Logarithm
  - Reflect and Inverse

Transformations



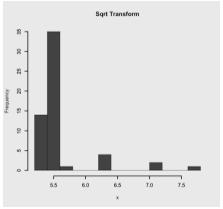
Positive skew

Skew = 3.05
Kurtosis = 9.26



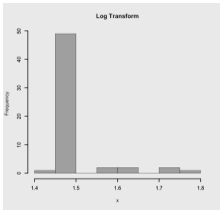
Square root transform

Skew = 2.86
Kurtosis = 7.91



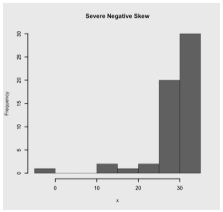
Log transform

Skew = 2.69
Kurtosis = 6.75



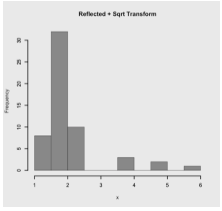
Negative skew

Skew = -3.36
Kurtosis = 11.73



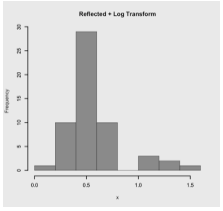
Reflect and square root

Skew = 2.58
Kurtosis = 6.62



Reflect and log

Skew = 1.59
Kurtosis = 2.92



Segment summary

- If a distribution is not normal then it is sometimes possible to transform the data in an attempt to make the distribution normal

Segment summary

- Most common transformations for positive skew
  - Square root
  - Logarithm
  - Inverse

Segment summary

- Most common transformations for negative skew
  - Reflect and Square root
  - Reflect and Logarithm
  - Reflect and Inverse

END SEGMENT

36

END LECTURE 21

37

Statistics One

Lecture 23
Generalized Linear Model

Two segments

- Overview
- Examples

Lecture 23 ~ Segment 1

Generalized Linear Model
Overview

Generalized Linear Model

- An extension of the General Linear Model that allows for non-normal distributions in the outcome variable and therefore also allows testing of non-linear relationships between a set of predictors and the outcome variable

Generalized Linear Model

- Generalized Linear Model: GLM*
- General Linear Model: GLM

5

General Linear Model (GLM)

- GLM is the mathematical framework used in many common statistical analyses, including multiple regression and ANOVA

6

Characteristics of GLM

- *Linear*: pairs of variables are assumed to have linear relations
- *Additive*: if one set of variables predict another variable, the effects are thought to be additive

7

Characteristics of GLM

- BUT! This does not preclude testing non-linear or non-additive effects

8

Characteristics of GLM

- GLM can accommodate such tests, for example, by
  - Transformation of variables
    - Transform so non-linear becomes linear
  - Moderation analysis
    - Fake the GLM into testing non-additive effects

9

GLM example

- Simple regression
  - $Y = B_0 + B_1X_1 + e$
  - Y = faculty salary
  - X1 = years since PhD

10

GLM example

- Multiple regression
  - $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + e$
  - Y = faculty salary
  - X1 = years since PhD
  - X2 = number of publications
  - X3 = (years x pubs)

11

Generalized linear model (GLM*)

- Appropriate when simple transformations or product terms are not sufficient

12

Generalized linear model (GLM*)

- The “linear” model is allowed to generalize to other forms by adding a “link function”

13

Generalized linear model (GLM*)

- For example, in binary logistic regression, the logit function was the link function

14

Binary logistic regression

- $\ln(\hat{Y} / (1 - \hat{Y})) = B_0 + \Sigma(B_k X_k)$

  $\hat{Y}$ = predicted value on the outcome variable Y
  B_0 = predicted value on Y when all X = 0
  X_k = predictor variables
  B_k = unstandardized regression coefficients
  $(Y – \hat{Y})$ = residual (prediction error)
  k = the number of predictor variables

15

Segment summary

- GLM* is an extension of GLM that allows for non-normal distributions in the outcome variable and therefore also allows testing of non-linear relationships between a set of predictors and the outcome variable

16

Segment summary

- Appropriate when simple transformations or product terms are not sufficient

17

Segment summary

- The “linear” model is allowed to generalize to other forms by adding a “link function”

18

END SEGMENT

Lecture 23 ~ Segment 2

Generalized Linear Model
Examples

GLM* Examples

- GLM* is an extension of GLM that allows for non-normal distributions in the outcome variable and therefore also allows testing of non-linear relationships between a set of predictors and the outcome variable

21

GLM* Examples

- Appropriate when simple transformations or product terms are not sufficient

22

GLM* Examples

- The “linear” model is allowed to generalize to other forms by adding a “link function”

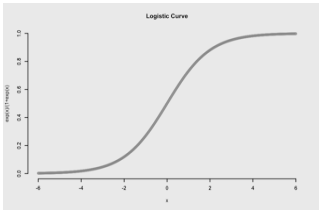
23

GLM* Examples

- Binary logistic regression

24

Binary logistic regression



GLM* Examples

- In binary logistic regression, the logit function served as the link function

26

GLM* Examples

- $\ln(\hat{Y} / (1 - \hat{Y})) = B_0 + \Sigma(B_k X_k)$

  $\hat{Y}$ = predicted value on the outcome variable Y
  B_0 = predicted value on Y when all X = 0
  X_k = predictor variables
  B_k = unstandardized regression coefficients
  $(Y - \hat{Y})$ = residual (prediction error)
  k = the number of predictor variables

27

GLM* Examples

- More than 2 categories on the outcome
  - Multinomial logistic regression
    - A-1 logistic regression equations are formed
      - Where A = # of groups
      - One group serves as reference group

GLM* Examples

- Another example is Poisson regression
- Poission distributions are common with "count data"
  - A number of events occurring in a fixed interval of time

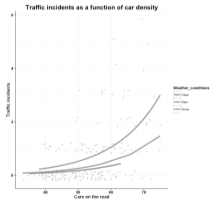
29

GLM* Examples

- For example, the number of traffic accidents as a function of weather conditions
  - Clear weather
  - Rain
  - Snow

30

Poisson regression



31

GLM* Examples

- In Poisson regression, the log function serves as the link function
- Note: this example also has a categorical predictor and would therefore also require dummy coding

32

Segment summary

- GLM* is an extension of GLM that allows for non-normal distributions in the outcome variable and therefore also allows testing of non-linear relationships between a set of predictors and the outcome variable

33

Segment summary

- Appropriate when simple transformations or product terms are not sufficient

34

Segment summary

- The “linear” model is allowed to generalize to other forms by adding a “link function”

35

END SEGMENT

END LECTURE 23

Statistics One

Lecture 24
Course Summary

1

Four segments

- Research methods and descriptive statistics
  - Lectures 1 – 6
- Simple and multiple regression
  - Lectures 7 – 14

2

Four segments

- Group comparisons with t-tests and ANOVA
  - Lectures 15 – 18
- Procedures for non-normal distributions and non-linear models
  - Lectures 19 – 23

3

Lecture 24 ~ Segment 1

Research Methods and
Descriptive Statistics

4

Research methods

- Descriptive research
- Experimental research
- Correlational research

Descriptive statistics

- Histograms
- Summary statistics
  - Measures of central tendency
    - Mean
    - Median
    - Mode
  - Measures of variability
    - Standard deviation
    - Variance

Descriptive statistics

- Correlation
- Covariance
- Scatterplots

Descriptive statistics

- Measurement
  - Classical true score theory
  - Reliability
  - Validity

END SEGMENT

9

Lecture 24 ~ Segment 2

Simple and multiple regression

10

Simple and multiple regression

- Simple regression equation has only one predictor variable (X)
- Multiple regression equation has multiple predictor variables

NHST

- NHST can be used to test statistical significance of individual predictor variables and to test statistical significance of the model

NHST

- Sampling
- Sampling error
- Sampling distribution
- Central limit theorem
- Problems with NHST
- Remedies

NHST

- Problems with NHST
  - BAYES
  - Biased by sample size
  - Arbitrary decision rule
  - Yokel local test
  - Error prone
  - Shady logic

NHST

- Remedies
  - Effect size
  - Confidence intervals
  - Model comparison
  - Replications
  - Power

Simple regression

- Regression equation
- Regression constant
- Regression coefficient (unstandardized and standardized)
- Residual

- Ordinary Least Squares

Mutiple regression

- Matrix algebra
- Regression equation (model)
- Regression constant
- Regression coefficients (unstandardized and standardized)
- Residual
- Model comparison

- Ordinary Least Squares

Mutiple regression

- Moderation
  - Dummy coding
  - Centering
- Mediation
  - Sobel test

END SEGMENT

19

Lecture 24 ~ Segment 3

Group Comparisons
t-tests and ANOVA

20

Group comparisons

- z-test
- Single sample t-test
- Independent t-test
  - Homogeneity of variance assumption
  - Levene's test
- Dependent t-test (paired samples)

Group comparisons

- ANOVA: One-way between groups
  - $F = MS_A = MS_{S/A}$
  - Homogeneity of variance assumption
  - Levene's test
  - Post-hoc tests

Group comparisons

- Factorial ANOVA
  - Main effects
  - Interaction effect
  - Simple effects
    - Homogeneity of variance assumption
    - Levene's test
    - Post-hoc tests

Group comparisons

- Repeated measures ANOVA
  - $F = MS_A = MS_{AxS}$
  - Sphericity assumption
  - Mauchly's test
  - Post-hoc tests

END SEGMENT

25

Lecture 24 ~ Segment 4

Procedures for non-normal distributions and non-linear models

26

Categorical outcome variables

- Chi-square tests
- Logistic regression

Non-normal distributions

- How to detect non-normal distributions
  - Histograms and scatterplots
  - Q-Q plots
- Common transformations
  - Square root
  - Logarithmic
  - Inverse

Non-parametric statistics

- Wilcoxan's ranking method
- Mann-Whitney U

Non-linear models

- Generalized Linear Model
  - Binomial
  - Multinomial
  - Poisson

END SEGMENT

31

END LECTURE 24

32