Stat 1600

Loren Heun

WMU

July 26, 2018

Statistics and Data Analysis

Lecture 2 Knowledge and data

Outline

Data Presentation # 1

- Statistics and Data
- Variable Types
- Summarizing Categorical Data

Knowledge and data

- Step-by-step knowledge building
- Some fallacies in interpreting evidence

Building knowledge step-by-step

1. Conceptualize the problem
   - This is the problem of interest. State it broadly.

Building knowledge step-by-step

2. Operationalize the problem
    - The investigator formulates specific questions to answer
    - What do you want to measure? These will become our dependent variables.

Building knowledge step-by-step

3. Design the Study
   - How will you select your sample?
   - How many groups will you compare?

Building knowledge step-by-step

4. Collect the Data
   - What instrument or technique will you use to collect data?
   - Will you use a survey, questionnaire, interview, observation?
   - Are you measuring a variable that will require special equipment/technology?

Building knowledge step-by-step

5. Analyze the Data
   - Are you comparing means? Percentages?
   - Are differences statistically significant?

Building knowledge step-by-step

6. Conclusions
   - Do the results generalize to a larger population?
   - Did you show cause-and-effect or just associations?

Building knowledge step-by-step

7. Disseminate results
    - How are you sharing your results?
    - News reports?
    - Scientific journals?
    - Etc. . .

Step-by-step Knowledge Building

- Conceptualize the problem – broad wording
- Operationalize the problem – specific questions
- Design the Study – how to select samples
- Collect the Data – measurement instrument
- Analyze the Data – comparing what
- Conclusions – repeatability and generalization
- Disseminate results – presentation of results

Example – comparing wt loss program

- Zone
  - Balances carbohydrates, protein, fat
- Atkins
  - Low carbohydrate, high fat, unrestricted calories
- LEARN
  - Low fat, and based on national guidelines
- Ornish
  - Low fat, high carbohydrate, unrestricted calories

How are we going to design a study to compare these?

Building knowledge step-by-step

1. Conceptualize the problem
   - Which weight loss program is most effective?
   - Which one is most healthy?

Building knowledge step-by-step

2. Operationalize the problem
   - How do we measure ‘effective’ and ‘healthy?’
   - At what time point are we interested in measuring? In 2 weeks, 2 months, 2 years?
   - Are we comparing average weight loss or perhaps the percentage of people who lost 15 pounds or more?
   - How do we measure healthy? LDL cholesterol reduction, BP reduction, Glucose levels?

Building knowledge step-by-step

3. Design the Study
   - Where are we recruiting our subjects?
   - How long will the study last?
   - Do they choose the diet or do we randomly assign them to it?
   - How do we ensure they stay on a diet?
   - What do we do with participants who go off the diet, do we eliminate them from the study?

Building knowledge step-by-step

4. Collect the Data
    - How many times will we measure their weights?
    - Are we taking blood samples? Urine samples? Are we sending samples to the lab?

Building knowledge step-by-step

5. Analyze the Data
   - Are there significant differences in average weight loss between the diet groups?
   - Are there differences in cholesterol, blood pressure, glucose levels or other biochemistry measures relating to health?
   - Are there differences in how well participants adhere to each diet plan?

Building knowledge step-by-step

6. Conclusions
    - After analyzing the results what do we conclude is the best diet? Why?
    - Can we generalize results to the larger population?
    - Are we sure weight loss can be attributed to the diet?

Building knowledge step-by-step

7. Disseminate results
   - How are we going to present the results?
   - What tables and graphs would make the study easy to read and understand?

Questioning results of a study

If we are reading the results of a study we need to be able to ask ourselves some questions:

- What is the long-term result (perhaps the results will differ if measurements are taken at longer time points)?
- What was the sample and to what population are we trying to generalize the results (males, females, age, ethnic differences)? We want to make sure we can generalize to the population outside of the study sample.
- Was the sample size large enough to allow for generalizing to the outside population?

Questioning results of a study

There is variation in study design, and some studies are designed better than others. We need to be able to judge the validity and reliability of a study.

Fallacies in interpreting evidence

1. Lack of evidence
   - "No proof that the drug is unsafe."
   - This is flawed as a lack of evidence does not mean the contrary is true and that the drug is safe.
2. Anecdotal evidence
   - "Testimonies of real people this worked for ..."
   - Infomercials.
   - Existence does not mean prevalence. Perhaps the drug or supplement worked for some people, but does that mean it is effective for the broader population?

Fallacies in interpreting evidence

- Correlation equals causation
  - "married people are happier than single people."
  - Did marriage cause the 'happier' outcome? Maybe happy people are the ones who tend to get married.
  - Two things happening at the same time does not mean one causes the other.

Examples of Wrong Reasoning Leading to Wrong Conclusions

- Lack of evidence fallacy. The fallacy lies in the reasoning that lack of evidence means the contrary is true.
- Anecdotal evidence fallacy. The fallacy lies in the reasoning that existence means prevalence.
- Correlation equals causation fallacy. The fallacy lies in the reasoning that "two things happening together" must mean one causes the other.

Statistics and Data Analysis

Ch 2.4 Summarizing Numerical Data

Outline

Data Presentation #2

- Summarizing Numerical Data

Sorted Data List

Payment (Rent or Mortgage), ACS Data:
140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

Sorted Data List

Payment (Rent or Mortgage), ACS Data:
140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

MIN = smallest observation = 140

Sorted Data List

Payment (Rent or Mortgage), ACS Data:
140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

MIN = smallest observation = 140

typical payment = around 700

Sorted Data List

Payment (Rent or Mortgage), ACS Data:
140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

MIN = smallest observation = 140
MAX = largest observation = 5200, an outlier
typical payment = around 700

Outlier

An observation that falls apart from the rest of the data
$\Rightarrow$ check for correctness

Here,

Household	State	Bedrooms	Payment	Type	Income
28	Michigan	4	5200	Mortgage	358000

$\Rightarrow$ OK

Stem-and-Leaf Plot

- See next slide and page 24 for two views of the payment data (Mortgage & Rent combined) using stem-and-leaf plots
- See page 25 for the comparison of the payments of the two types, Mortgage and Rent, using side-by-side stem-and-leaf plots (same scale, i.e., same stem width)

Stem-and-Leaf Plot

```
The decimal point is 2 digit(s) to the right of the |

 1 | 49
 2 | 0023589
 3 | 445788
 4 | 02459
 5 | 0001356
 6 | 577
 7 | 00001244567
 8 | 058
 9 | 00179
10 | 00
11 | 0
12 | 00000
13 | 0
14 | 00
15 | 0
16 |
17 |
18 | 0
```

Note: 140, the one (1) to the left of | is the hundredths digit and the four (4) to the right of | is the tens digit.

Stem-and-Leaf Plot

```
  The decimal point is 2 digit(s) to the right of the |

   1 | 49
   2 | 0023589
   3 | 445788
   4 | 02459
   5 | 0001356
   6 | 577
   7 | 00001244567
   8 | 058
   9 | 00179
  10 | 00
  11 | 0
  12 | 00000
  13 | 0
  14 | 00
  15 | 0
  16 |
  17 |
  18 | 0
```

Note: 190, the one (1) to the left of | is the hundredths digit and the nine (9) to the right of | is the tens digit.

iClicker Question 2.4.1

The stem-and-leaf display below shows the BMI (Body Mass Index) of 14 individuals. What number(s) does ‘2 | 56’ represent?

1. 2.5 and 2.6
2. 25 and 26
3. 250 and 260
4. 256
5. None of the above

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

1 | 588
2 | 11222334
2 | 56
3 | 1
```

Relative Frequency Table

The data range is first divided into several (usually) equal-width class intervals and then we obtain the frequency/relative frequency of data values contained in each class interval.

- Often has 5 to 15 intervals (depending on number of observations)
- Starting value of the first (i,e, left-most) interval = _____.
- Settle boundary disputes (for example we may have intervals contain the left endpoint but not the right)

Relative Frequency Table

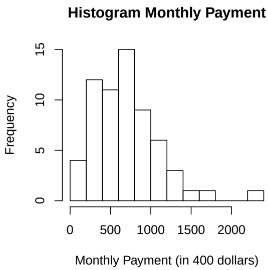
Monthly Payment($)*	Frequency	Rel. freq.(%)
0-200	2	3.2
200-400	13	20.6
400-600	12	19.0
600-800	14	22.2
800-1000	8	12.7
1000-1200	3	4.8
1200-1400	6	9.5
1400-1600	3	4.8
1600-1800	0	0
1800-2000	1	1.6
2000-2200	0	0
2200-2400	0	0
2400-2600	1	1.6
Total	63	100

* Interval contain the left endpoint, but not the right

Histogram

A graphical display of the relative frequency table defined by the class intervals

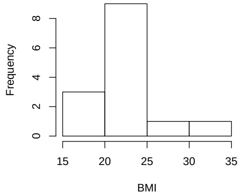
$\Rightarrow$ frequencies (or relative frequencies) are plotted as columns



iClicker Question 2.4.2

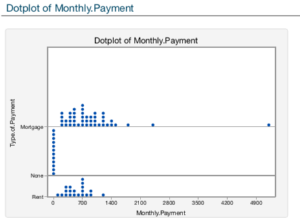
The histogram below shows the BMI (Body Mass Index) of 15 individuals. The right inclusion rule was used in the construction of the histogram. What class interval(s) occurs least frequently?

1. (25, 30]
2. (30, 35]
3. (20, 25]
4. (25, 30] and (30, 35]
5. Unknown



Dotplot

- Each observation is represented by a dot, repeated values are stacked upwards
- Below is a comparison dotplot of the monthly payments from the two types of payment:



iClicker Question 2.4.3

We summarize numerical data with all of the following EXCEPT:

1. Bar chart
2. Dotplot
3. Histogram
4. Scatterplot
5. Stem and leaf

Statistics and Data Analysis

STAT 1600
Ch 2.4.4 Box and Whisker Plot

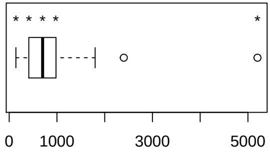
Outline

Summarizing Numerical Data, #2

- Box-and-Whisker Plot
- Symmetry and Skewness

Box -and-whisker plot (or Boxplot)

This is a graphical picture of the distribution of quarters of the data



- **This shows the range of each of the four quarters of data:**
- MINimum to Q1 (upper boundary of first quarter): 140, 415
- Median (upper boundary of second quarter, also known as Q2): 700
- Q3 is the upper boundary of the third quarter: 975
- MAXimum is the largest of the ordered observations: 5200

Five-number Summary

Payment (Rent/Mortgage, outlier excluded), ACS Data
$n = 63$; $(n+1)/4 = 16$; 1st quarter of the data

140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

The range is from 140 to 415

Five-number Summary

Payment (Rent/Mortgage, outlier excluded), ACS Data
$n = 63$; $(n + 1)/4 = 16$; 2nd quarter of the data

140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

The range is from 415 to 700

Five-number Summary

Payment (Rent/Mortgage, outlier excluded), ACS Data
$n = 63$; $(n + 1)/4 = 16$; 3rd quarter of the data

140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

The range is from 700 to 975

Five-number Summary

Payment (Rent/Mortgage, outlier excluded), ACS Data
$n = 63$; $(n + 1)/4 = 16$; 4th quarter of the data

140, 190, 200, 200, 220, 230, 250, 280, 290, 340, 340, 350, 370, 380, 380, 400, 420, 440, 450, 490, 500, 500, 500, 510, 530, 550, 560, 650, 670, 670, 700, 700, 700, 700, 710, 720, 740, 740, 750, 760, 770, 800, 850, 880, 900, 900, 910, 970, 990, 1000, 1000, 1100, 1200, 1200, 1200, 1200, 1200, 1300, 1400, 1400, 1500, 1800, 2400, 5200

The range is from 975 to 2400

Five-number Summary

Payment (Rent/Mortgage, outlier excluded), ACS Data
$n = 63$; $(n + 1)/4 = 16$; In summary,

- MIN = smallest observation = 140
- Q1 = 1st quartile = 400
- MED = median (= 2nd quartile) = 700
- Q3 = 3rd quartile = 970
- MAX = largest observation = 2400
-
- MIN, Q1, MED, Q3, and MAX give five-number summary

Five-number Summary

Payment (Rent/Mortgage, outlier excluded), ACS Data
$n = 63$; $(n + 1)/4 = 16$; In summary,

- 50% of data values $\leq$ MED
- 50% of data values $\geq$ MED
- 25% of data values $\leq$ Q1
- 75% of data values $\geq$ Q1

Computing Five-number Summary

1. Sort data into a list of ordered values

Computing Five-number Summary

1. Sort data into a list of ordered values
2. Find MIN and MAX

Computing Five-number Summary

1. Sort data into a list of ordered values
2. Find MIN and MAX
3. Determine the sample size (i.e., number of observations) n

Computing Five-number Summary

1. Sort data into a list of ordered values
2. Find MIN and MAX
3. Determine the sample size (i.e., number of observations) n
4. Q1 $= 0.25(n + 1)$th ordered value

Computing Five-number Summary

1. Sort data into a list of ordered values
2. Find MIN and MAX
3. Determine the sample size (i.e., number of observations) n
4. $Q1 = 0.25(n + 1)$th ordered value
5. $MED = 0.5(n + 1)$th ordered value

Computing Five-number Summary

1. Sort data into a list of ordered values
2. Find MIN and MAX
3. Determine the sample size (i.e., number of observations) n
4. Q1 = 0.25(n + 1)th ordered value
5. MED = 0.5(n + 1)th ordered value
6. Q3 = 0.75(n + 1)th ordered value

Computing Five-number Summary

1. Sort data into a list of ordered values
2. Find MIN and MAX
3. Determine the sample size (i.e., number of observations) n
4. Q1 $= 0.25(n + 1)$th ordered value
5. MED $= 0.5(n + 1)$th ordered value
6. Q3 $= 0.75(n + 1)$th ordered value
7. If a non-integer resulted in any computation of the quartiles (Q1, MED, Q3) above, average the two adjacent ordered values for the respective quartile

Five-number Summary for Monthly Rent

Sorted monthly payments for Type = Rent ($n = 20$), ACS Data

140, 220, 250, 280, 340, 350, 380, 400, 490, 500, 560, 650, 670, 700, 700, 740, 760, 880, 910, 1200

1. MIN = 140, MAX = 1200

Five-number Summary for Monthly Rent

Sorted monthly payments for Type = Rent ($n = 20$), ACS Data

140, 220, 250, 280, 340, 350, 380, 400, 490, 500, 560, 650, 670, 700, 700, 740, 760, 880, 910, 1200

1. MIN = 140, MAX = 1200
2. $0.25(n + 1) = 5.25$ and hence Q1 = average of 5th and 6th ordered values = $(340 + 350)/2 = 345$

Five-number Summary for Monthly Rent

Sorted monthly payments for Type = Rent ($n = 20$), ACS Data

140, 220, 250, 280, 340, 350, 380, 400, 490, 500, 560, 650, 670, 700, 700, 740, 760, 880, 910, 1200

1. MIN = 140, MAX = 1200
2. $0.25(n + 1) = 5.25$ and hence Q1 = average of 5th and 6th ordered values = $(340 + 350)/2 = 345$
3. $0.5(n + 1) = 10.5$ and hence MED = average of 10th and 11th ordered values = $(500 + 560)/2 = 530$

Five-number Summary for Monthly Rent

Sorted monthly payments for Type = Rent ($n = 20$), ACS Data

140, 220, 250, 280, 340, 350, 380, 400, 490, 500, 560, 650, 670, 700, 700, 740, 760, 880, 910, 1200

1. MIN = 140, MAX = 1200
2. $0.25(n + 1) = 5.25$ and hence Q1 = average of 5th and 6th ordered values $= (340 + 350)/2 = 345$
3. $0.5(n + 1) = 10.5$ and hence MED = average of 10th and 11th ordered values $= (500 + 560)/2 = 530$
4. $0.75(n + 1) = 15.75$ and hence Q3 = average of 15th and 16th ordered values $= (700 + 740)/2 = 720$

iClicker Question 2.4.4.1

In general, the middle 50% of data values are bounded by what statistics?

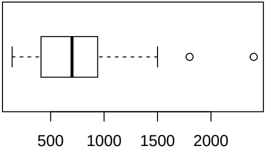
1. The first quartile and the median.
2. The first quartile and the third quartile.
3. The median and the third quartile.
4. The median and the maximum.
5. The minimum and the maximum.

iClicker Question 2.4.4.2

Given the 5-Number summary (MIN, Q1, Q2, Q3, and MAX) of any data set, approximately 75% of data values are at or above what statistic?

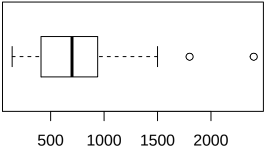
1. The median.
2. The first quartile.
3. The third quartile.
4. The maximum.
5. The minimum.

Box-and-whisker Plot



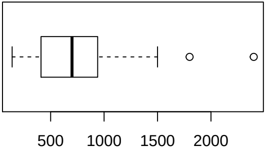
- Draw a horizontal axis covering data range

Box-and-whisker Plot



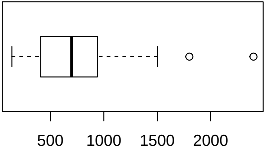
- Draw a horizontal axis covering data range
- Draw a box with edges at Q1 and Q3

Box-and-whisker Plot



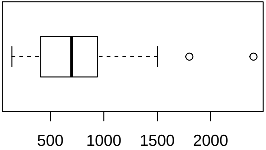
- Draw a horizontal axis covering data range
- Draw a box with edges at Q1 and Q3
- Draw within the box, a line located at MED

Box-and-whisker Plot



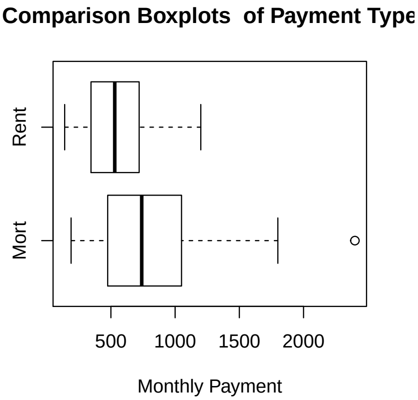
- Draw a horizontal axis covering data range
- Draw a box with edges at Q1 and Q3
- Draw within the box, a line located at MED
- Draw ‘fences’ (lines) at the MIN and MAX

Box-and-whisker Plot



- Draw a horizontal axis covering data range
- Draw a box with edges at Q1 and Q3
- Draw within the box, a line located at MED
- Draw 'fences' (lines) at the MIN and MAX
- Draw 'whiskers' extending from the edges of the box to the MIN and MAX

Comparison Boxplots



Symmetry/Skewness

Left-skewed

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

  4 | 0
  5 | 5
  6 | 2
  7 |
  8 | 17
  9 | 2445
 10 | 011245569
 11 | 58
```

Symmetry/Skewness

Symmetric

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

   4 | 7
   5 | 35
   6 | 005
   7 | 00112
   8 | 467899
   9 | 0199
  10 | 14
  11 | 1
```

Symmetry/Skewness

Right-skewed

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

  4 | 58
  5 | 011245569
  6 | 2445
  7 | 17
  8 |
  9 | 2
 10 | 5
 11 | 0
```

Symmetry/Skewness, continued

- **Symmetric:** data shape in two mirror-imaged halves

Symmetry/Skewness, continued

- **Symmetric:** data shape in two mirror-imaged halves
- **Right-skewed:** long right tail

Symmetry/Skewness, continued

- **Symmetric:** data shape in two mirror-imaged halves
- **Right-skewed:** long right tail
- **Left-skewed:** long left tail

Symmetry/Skewness, continued

- **Symmetric:** data shape in two mirror-imaged halves
- **Right-skewed:** long right tail
- **Left-skewed:** long left tail
- **Symmetry/Skewness** can be detected by inspecting the stem-and-leaf plot (first turn it counter-clockwise 90 degrees), histogram, dotplot, or boxplot (median is half way from the edges of the box, whiskers on two sides of equal length)

iClicker Question 2.4.4.3

The stem-and-leaf displays of two data sets are given below. Describe the shape of these two data sets.

1. 1. symmetric; 2. left skewed
2. 1. right skewed; 2. symmetric
3. 1. symmetric; 2. symmetric
4. 1. left skewed; 2. left skewed
5. 1. symmetric; 2. right skewed

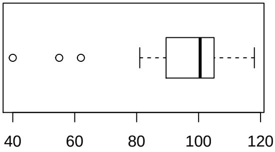
	Set 1
1	599
2	0113
3	669

	Set 2
0	1223466
1	0015
2	9

iClicker Question 2.4.4.4

Describe the shape of the box-and-whisker plot (boxplot) below.

1. Symmetric
2. Right skewed
3. Left skewed
4. right and left skewed
5. None of the above



Statistics and Data Analysis

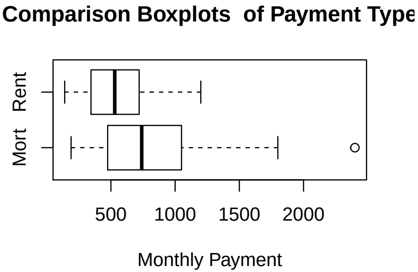
STAT 1600
Ch. 3 Estimates of Center

Outline

Summarizing Numerical Data, #3

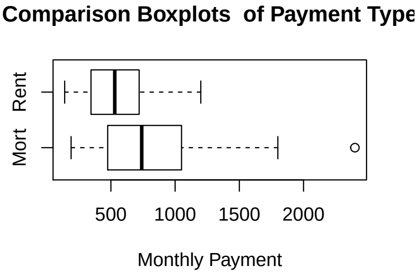
- Location and Spread

Importance of Location and Spread



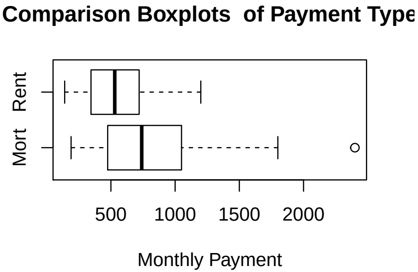
- The central *location* of Rent appears to be smaller than that of Mortgage

Importance of Location and Spread



- The central *location* of Rent appears to be smaller than that of Mortgage
- Moreover, the *spread* of Rent appears to be smaller than that of Mortgage, i.e., Rent payment is less scattered (or less variable) than that of Mortgage

Importance of Location and Spread



- The central *location* of Rent appears to be smaller than that of Mortgage
- Moreover, the *spread* of Rent appears to be smaller than that of Mortgage, i.e., Rent payment is less scattered (or less variable) than that of Mortgage
- But how do we quantify such comparisons? $\Rightarrow$ through **location** and **spread** measures

Outline

Summarizing Numerical Data, #3

- Location and spread
  - Location and spread
- Estimates of Center
  - Estimating Average Value
  - The Sample Means
  - The Trimmed Mean
  - The Median of Pairwise Averages

How do you estimate the average rental

Based on the following random sample of 2-bedroom apartments in Kalamazoo area

140, 220, 280, 1200, 380, 560, 400, 880, 700, 670

How do you measure the (central) location of such rentals?

Sample Mean is an Est. of Pop. Mean

- The sample mean (i.e., average of the sample) is denoted by $\bar{X}$

Sample Mean is an Est. of Pop. Mean

- The sample mean (i.e., average of the sample) is denoted by $\bar{X}$
- and is the arithmatic average of data values i.e.,

Sample Mean is an Est. of Pop. Mean

- The sample mean (i.e., average of the sample) is denoted by $\bar{X}$
- and is the arithmatic average of data values i.e.,
- $$\bar{X} = \frac{\texttt{sum of data values}}{\texttt{sample size}}$$

Sample Mean is an Est. of Pop. Mean

- The sample mean (i.e., average of the sample) is denoted by $\bar{X}$
- and is the arithmatic average of data values i.e.,
- $$\bar{X} = \frac{\texttt{sum of data values}}{\texttt{sample size}}$$
- (2-bedroom appartment rental example)

Sample Mean is an Est. of Pop. Mean

- The sample mean (i.e., average of the sample) is denoted by $\bar{X}$
- and is the arithmatic average of data values i.e.,
- $$\bar{X} = \frac{\texttt{sum of data values}}{\texttt{sample size}}$$
- (2-bedroom appartment rental example)
- $$\bar{X} = \frac{140 + 220 + \cdots + 670}{10} = 543$$

Sample Mean is an Est. of Pop. Mean

- That is, the average rent for 2-bedroom apartments in Kalamazoo area is \$ 543 in our sample

Sample Mean is an Est. of Pop. Mean

- That is, the average rent for 2-bedroom apartments in Kalamazoo area is \$ 543 in our sample
- This is not to be interpreted as the actual *population average* (i.e., the actual average rent of all 2-bedroom apartments in the entire Kalamazoo area)

Sample Mean is an Est. of Pop. Mean

- That is, the average rent for 2-bedroom apartments in Kalamazoo area is $ 543 in our sample
- This is not to be interpreted as the actual *population average* (i.e., the actual average rent of all 2-bedroom apartments in the entire Kalamazoo area)
- It is subject to *sampling error*

Sample Mean is an Est. of Pop. Mean

- That is, the average rent for 2-bedroom apartments in Kalamazoo area is \$ 543 in our sample
- This is not to be interpreted as the actual *population average* (i.e., the actual average rent of all 2-bedroom apartments in the entire Kalamazoo area)
- It is subject to *sampling error*
- It likely missed the true population mean (denoted μ), by $|\bar{X} - \mu|$, the *sampling error*.

The Sample Median

- an alternative to the sample mean as a measure of location
- Recall that the median is the $0.5(n + 1)$th ordered data value
- Sorted list of Kalamazoo 2-bedroom rental data

  140, 220, 280, 380, 400, 560, 670, 700, 880, 1200

The Sample Median

- an alternative to the sample mean as a measure of location
- Recall that the median is the 0.5(n + 1)th ordered data value
- Sorted list of Kalamazoo 2-bedroom rental data

  140, 220, 280, 380, 400, 560, 670, 700, 880, 1200

- $$0.5(n+1) = 0.5 \times 11 = 5.5$$

The Sample Median

- an alternative to the sample mean as a measure of location
- Recall that the median is the 0.5(n + 1)th ordered data value
- Sorted list of Kalamazoo 2-bedroom rental data

  140, 220, 280, 380, 400, 560, 670, 700, 880, 1200

- $$0.5(n+1) = 0.5 \times 11 = 5.5$$
- 5.5 is in-between 5 and 6

The Sample Median

- an alternative to the sample mean as a measure of location
- Recall that the median is the 0.5(n + 1)th ordered data value
- Sorted list of Kalamazoo 2-bedroom rental data

  140, 220, 280, 380, 400, 560, 670, 700, 880, 1200

- $$0.5(n+1) = 0.5 \times 11 = 5.5$$
- 5.5 is in-between 5 and 6
- Hence, MED = average of 5th & 6th ordered values

The Sample Median

- an alternative to the sample mean as a measure of location
- Recall that the median is the 0.5(n + 1)th ordered data value
- Sorted list of Kalamazoo 2-bedroom rental data

  140, 220, 280, 380, 400, 560, 670, 700, 880, 1200

- $$0.5(n+1) = 0.5 \times 11 = 5.5$$
- 5.5 is in-between 5 and 6
- Hence, MED = average of 5th & 6th ordered values
- $$\tilde{x} = \frac{400+560}{2} = 480$$

iClicker Question 3.1.1

The fuel efficiency (MPG) of 5 Japanese made cars are listed below

$$27.5, 27.2, 34.1, 29.5, 31.8$$

What is the median MPG?

1. 29.5
2. 34.1
3. 30.50
4. 28.5
5. 27.5

Sample Mean versus Sample Median

- Sample mean is sensitive to outliers

Sample Mean versus Sample Median

- Sample mean is sensitive to outliers
- Sample median is *insensitive* to outliers

Sample Mean versus Sample Median

- Sample mean is sensitive to outliers
- Sample median is *insensitive* to outliers
- In the Kalamazoo apartment rental data, what if the smallest value \$ 140 is replaced by \$100?

Sample Mean versus Sample Median

- Sample mean is sensitive to outliers
- Sample median is *insensitive* to outliers
- In the Kalamazoo apartment rental data, what if the smallest value \$ 140 is replaced by \$100?
- The MED remains unchanged (\$ 480) but $\bar{X} = 539$, down by from the original data (\$ 543).

Sample Mean versus Sample Median

- Looking at another example:
- Let's say we have a dataset of the following:
- Data: 5, 10, 17, 20, 25
- Mean: **15.4**; Median:17
- Data: 5, 10, 17, 20, 40
- Mean: **18.4**; Median: 17

We can see the median has not been affected by the outlier, whereas the mean has been affected.

The Trimmed Mean

- The Trimmed Mean is less sensitive to outliers when compared with sample mean

The Trimmed Mean

- The Trimmed Mean is less sensitive to outliers when compared with sample mean
- 10% trimmed mean = mean of data with lowest 10% values and highest 10% values excluded = mean of middle 80% values

The Trimmed Mean

- The Trimmed Mean is less sensitive to outliers when compared with sample mean
- 10% trimmed mean = mean of data with lowest 10% values and highest 10% values excluded = mean of middle 80% values
- in Kalamazoo apartment rental data, 10% of n = 10 is 1 $(n \times 0.1 = 1)$

  220, 280, 380, 400, 560, 670, 700, 880

The Trimmed Mean

- The Trimmed Mean is less sensitive to outliers when compared with sample mean
- 10% trimmed mean = mean of data with lowest 10% values and highest 10% values excluded = mean of middle 80% values
- in Kalamazoo apartment rental data, 10% of n = 10 is 1 $(n \times 0.1 = 1)$

  220, 280, 380, 400, 560, 670, 700, 880
- Hence, 10% trimmed mean, denoted $\bar{X}_{tr}$, is computed by

$$\begin{aligned}\bar{X}_{tr} &= \frac{220 + 280 + 380 + 400 + 560 + 670 + 700 + 880}{8} \\ &= 511.25\end{aligned}$$

The Trimmed Mean – cont'd

- If $n \times 0.1$ is not an integer, round it up. E.g., if $n = 23$ such that $n \times 0.1 = 2.3$ then exclude the 3 lowest values and the 3 highest values, thus computing the trimmed mean as the average of 17 $(= 23 - 3 - 3)$ middle values to ensure at least 10% protection (against outlying values at each end)

The Trimmed Mean – cont'd

- If $n \times 0.1$ is not an integer, round it up. E.g., if $n = 23$ such that $n \times 0.1 = 2.3$ then exclude the 3 lowest values and the 3 highest values, thus computing the trimmed mean as the average of 17 $(= 23 - 3 - 3)$ middle values to ensure at least 10% protection (against outlying values at each end)
- The dataset must be ordered.

The Median of Pairwise Averages

- The median of pairwise averages is another compromise between the mean and the median.
- We replace observations by pairwise averages of those observations.
- Next we take the median of those.
- Also make sure to pair each observation with itself!
- Proceed in the following pattern as laid out on the next slide.

The Median of Pairwise Averages

$\frac{500+500}{2}$	$\frac{500+500}{2}$	$\frac{500+525}{2}$	$\frac{500+555}{2}$	$\frac{500+635}{2}$	$\frac{500+650}{2}$	$\frac{500+670}{2}$	$\frac{500+675}{2}$	$\frac{500+750}{2}$	$\frac{500+800}{2}$
	$\frac{500+500}{2}$	$\frac{500+525}{2}$	$\frac{500+555}{2}$	$\frac{500+635}{2}$	$\frac{500+650}{2}$	$\frac{500+670}{2}$	$\frac{500+675}{2}$	$\frac{500+750}{2}$	$\frac{500+800}{2}$
		$\frac{525+525}{2}$	$\frac{525+555}{2}$	$\frac{525+635}{2}$	$\frac{525+650}{2}$	$\frac{525+670}{2}$	$\frac{525+675}{2}$	$\frac{525+750}{2}$	$\frac{525+800}{2}$
			$\frac{555+555}{2}$	$\frac{555+635}{2}$	$\frac{555+650}{2}$	$\frac{555+670}{2}$	$\frac{555+675}{2}$	$\frac{555+750}{2}$	$\frac{555+800}{2}$
				$\frac{635+635}{2}$	$\frac{635+650}{2}$	$\frac{635+670}{2}$	$\frac{635+675}{2}$	$\frac{635+750}{2}$	$\frac{635+800}{2}$
					$\frac{650+650}{2}$	$\frac{650+670}{2}$	$\frac{650+675}{2}$	$\frac{650+750}{2}$	$\frac{650+800}{2}$
						$\frac{670+670}{2}$	$\frac{670+675}{2}$	$\frac{670+750}{2}$	$\frac{670+800}{2}$
							$\frac{675+675}{2}$	$\frac{675+750}{2}$	$\frac{675+800}{2}$
								$\frac{750+750}{2}$	$\frac{750+800}{2}$
									$\frac{800+800}{2}$

- averaging 1st obs with itself = (500 + 500) / 2 = 500,
- averaging 1st obs with 2nd obs = (500 + 500) / 2 = 500, and so on ...
- averaging 2nd obs with itself = (500 + 500) / 2 = 500, and so on ...
- and so on ...
- median of 55 (= .5(n + 1)/2) pairwise averages = 28th (0.5(55 + 1) = 28) ordered pairwise average = $625.0.

Robustness of Est. of Central Location

- Recall that an estimate is robust if it is insensitive to outliers.
- **Robust** = resistant to errors
- The sample mean is NOT robust. That is, it is sensitive to outliers.
- The sample median and the median of pairwise averages are robust.
- The trimmed means are more robust than the mean but less robust than the median. Trimmed means with higher trimmed percentage are more robust.

Clicker Question 3.1.2

For estimates of central location, which of the following statements is true?

1. Median, mean, and the median of pairwise averages are robust.
2. Median, 20% trimmed mean, and mean are robust.
3. Mean and the median are not robust.
4. Median and the median of pairwise averages are robust.
5. All statements above are incorrect.

Statistics and Data Analysis

STAT 1600
Ch 3.2 Estimates of Spread

Outline

Estimates of Spread

- Estimates of Spread
- The Sample Standard Deviation
- Effect of Multiplication/Addition by a Constant

Estimates of Spread

(or Uncertainty, Variation)

- An estimate of spread is a measure of uncertainty, or variation, or 'give or take'

Estimates of Spread

(or Uncertainty, Variation)

- An estimate of spread is a measure of uncertainty, or variation, or 'give or take'
- When two or more comparable data sets (comparable means data sets are of same type/same unit of numerical measurements) are compared, the one with the smallest spread has the least uncertainty around the estimate of center (i.e., least scattered)

Estimates of Spread

(or Uncertainty, Variation)

- An estimate of spread is a measure of uncertainty, or variation, or 'give or take'
- When two or more comparable data sets (comparable means data sets are of same type/same unit of numerical measurements) are compared, the one with the smallest spread has the least uncertainty around the estimate of center (i.e., least scattered)
- Estimates of spread are non-negative

Estimates of Spread, Cont'd

i.e.,

- The standard deviation (SD) is typically used as a 'give or take' number in describing the spread of a dataset

Estimates of Spread, Cont'd

i.e.,

- The standard deviation (SD) is typically used as a 'give or take' number in describing the spread of a dataset
-

$$\texttt{Sample SD} = s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} = \sqrt{\frac{SS}{n-1}}$$

The Sample Standard Deviation (SD)

	Rent	Diff	Diff Sqd
1	140	-403	162409
2	220	-323	104329
3	280	-263	69169
4	1200	657	431649
5	380	-163	26569
6	560	17	289
7	400	-143	20449
8	880	337	113569
9	700	157	24649
10	670	127	16129

- Step 1. Compute $\bar{x} = 543$

The Sample Standard Deviation (SD)

	Rent	Diff	Diff Sqd
1	140	-403	162409
2	220	-323	104329
3	280	-263	69169
4	1200	657	431649
5	380	-163	26569
6	560	17	289
7	400	-143	20449
8	880	337	113569
9	700	157	24649
10	670	127	16129

- Step 1. Compute $\bar{x} = 543$
- Step 2. Calculate Diff, i.e., how much an obs. 'missed by' the average, (i.e., deviation).

The Sample Standard Deviation (SD)

	Rent	Diff	Diff Sqd
1	140	-403	162409
2	220	-323	104329
3	280	-263	69169
4	1200	657	431649
5	380	-163	26569
6	560	17	289
7	400	-143	20449
8	880	337	113569
9	700	157	24649
10	670	127	16129

- Step 1. Compute $\bar{x} = 543$
- Step 2. Calculate Diff, i.e., how much an obs. ‘missed by’ the average, (i.e., deviation).
- Step 3. Square the ‘missed by’ differences.

The Sample Standard Deviation (SD)

	Rent	Diff	Diff Sqd
1	140	-403	162409
2	220	-323	104329
3	280	-263	69169
4	1200	657	431649
5	380	-163	26569
6	560	17	289
7	400	-143	20449
8	880	337	113569
9	700	157	24649
10	670	127	16129

- Step 1. Compute $\bar{x} = 543$
- Step 2. Calculate Diff, i.e., how much an obs. 'missed by' the average, (i.e., deviation).
- Step 3. Square the 'missed by' differences.
- Step 4. Add all the squared 'missed by' differences, i.e., SS

The Sample Standard Deviation (SD)

	Rent	Diff	Diff Sqd
1	140	-403	162409
2	220	-323	104329
3	280	-263	69169
4	1200	657	431649
5	380	-163	26569
6	560	17	289
7	400	-143	20449
8	880	337	113569
9	700	157	24649
10	670	127	16129

- Step 1. Compute $\bar{x} = 543$
- Step 2. Calculate Diff, i.e., how much an obs. 'missed by' the average, (i.e., deviation).
- Step 3. Square the 'missed by' differences.
- Step 4. Add all the squared 'missed by' differences, i.e., SS
- Step 5. Take the square-root of $\frac{SS}{n-1}$ to get SD

$$SD = \sqrt{\frac{9.6921 \times 10^5}{10 - 1}} = 328.2$$

Interpretation of SD

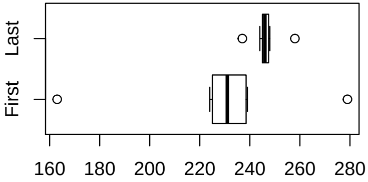
Table: Bowling Example

Games	Scores	Mean	SD
First games	163, 231, 224, 238, 279, 239, 226	228.6	34.34
Last games	246, 244, 247, 248, 237, 258, 246	246.6	6.21

Scores of Walter Ray Williams Jr. in 2008 bowling tournament, Indiana

Interpretation of SD

Scores of Walter Ray Williams Jr. in 2008 bowling tournament, Indiana



- Bigger swings (larger SD) in earlier games and scored typically lower (smaller Mean)
- He played consistently (smaller SD) in later games, and typically with better scores (larger Mean)

Sample Standard Deviation is Not Robust

As an estimate of the spread of a data set, the sample standard deviation **is sensitive to outliers**.

iClicker Question 3.2.1

The fuel efficiency (MPG) of 5 Japanese made cars are listed below

27.5, 27.2, 34.1, 29.5, 31.8

Ignoring any rounding error, what is the sum of all the deviations (MPG for Japanese made cars) from the mean MPG for Japanese made cars?

1. 9.38
2. 4.19
3. 0.00
4. -4.19
5. 30.58

iClicker Question 3.2.2

Recall that an estimate is robust if it is insensitive to outliers. Which of the following statements is true.

1. The sample mean and the standard deviation are robust.
2. The sample mean is robust but the standard deviation is not.
3. The sample mean is not robust but the standard deviation is.
4. The sample mean and the standard deviation are not robust.
5. None of the previous statements is true.

Effect of Multiplication/Addition by a Constant

apartment rental example

- Recall that the mean and SD are \$ 543 ± 328.2 ($\pm$ means 'give or take')

Effect of Multiplication/Addition by a Constant

apartment rental example

- Recall that the mean and SD are \$ 543 ± 328.2 ($\pm$ means 'give or take')
- Get a roommate and pay half the rent: \$ 271.5 ± 164.1

Effect of Multiplication/Addition by a Constant

apartment rental example

- Recall that the mean and SD are \$ 543 ± 328.2 ($\pm$ means 'give or take')
- Get a roommate and pay half the rent: \$ 271.5 ± 164.1
- No roommate but has contribution of \$100 per month from parents: 271.5 ± 328.2

General Rules

- when a constant is **added to/subtracted from** each data value, the same thing happens to the average, but the SD remains unchanged,

General Rules

- when a constant is **added to/subtracted from** each data value, the same thing happens to the average, but the SD remains unchanged,
- when each data value is **multiplied or divided** by a positive constant, the same thing happens to both the average and the SD.

General Rules

Eg. data: 1, 2, 3

X	Diff	Sqd
1	-1	1
2	0	0
3	1	1
$\bar{x} = 2$	0	$SS = 2$

$$SD = \sqrt{\frac{2}{(3-1)}} = 1$$

Eg. data: 3, 4, 5 (added 2)

X	Diff	Sqd
3	-1	1
4	0	0
5	1	1
$\bar{x} = 4$	0	$SS = 2$

$$SD = \sqrt{\frac{2}{(3-1)}} = 1$$

Notice here the mean increased from 2 to 4, yet the SD did not change.

General Rules

Eg. data: 1, 2, 3

X	Diff	Sqd
1	-1	1
2	0	0
3	1	1
$\bar{x} = 2$	0	$SS = 2$

$$SD = \sqrt{\frac{2}{(3-1)}} = 1$$

Eg. data: 2, 4, 6 (times 2)

X	Diff	Sqd
2	-2	4
4	0	0
6	2	4
$\bar{x} = 4$	0	$SS = 8$

$$SD = \sqrt{\frac{4}{(3-1)}} = 2$$

In this second example, multiplying by 2 the mean doubled AND the SD doubled.

iClicker Question 3.2.3

Compute the mean given the following data:
8, 12, 15, 22, 28

1. 5
2. 10
3. 15
4. 17
5. 20

Statistics and Data Analysis

STAT1600
Ch 4 Threats to Valid Comparisons

Outline

Threats to Valid Comparisons

- Hidden Confounder

Hidden Confounder

Lower Extremity Fractures Example
In the study of 'Lower extremity fractures in motor vehicle collisions: Influence of direction of impact and seatbelt use,' one of the conclusions is of interest: there was a higher incidence of lower extremity fracture among women.

Lower Extremity Fractures

Wrong assumptions may lead to wrong conclusions by falsely assuming that gender **causes** higher/lower leg fractures, one may reach these **false conclusions** about the study outcome ‘women have higher rates of leg fractures’

- because they drive faster?
- apply brakes more slowly?
- have weaker bones?

Lower Extremity Fractures

A follow-up study titled 'the role of driver gender and height' turns out that **height** was the culprit.
Because **height** and **gender** have a strong link, a false conclusion can result from a false assumption.
The pathway graph below describes the true relationship:

```
Confounder         Height  ⇒  Leg Fracture  Outcome
(actual cause)     ↕
Prob cause         Gender
```

↕ Means 'association' or 'probable cause'

⇒ Means 'cause-and-effect'

Lower Extremity Fractures

$$\texttt{pathway graph} : \begin{array}{lcl} \texttt{Height} & \Rightarrow & \texttt{Leg Fracture} \\ \updownarrow & & \\ \texttt{Gender} & & \end{array}$$

- outcome variable: leg fracture.
- probable cause: gender.
- confounder or confounding variable: the hidden variable height.

Confounding Variable or Confounder

A **confounder** is a third variable that is associated with both the probable cause and the **outcome**.

- *Overlooking* its existence can lead us to drawing a *wrong conclusion* about the cause-and-effect relationship.
- Hidden confounders are one of the biggest sources of (often unknowingly) *invalid comparisons*. One could end up comparing apples to oranges! (Note: women as a group are shorter than men!)

Lower Extremity Fractures, cont'd

Potential confounders other than height includes

- weight,
- none/light/heavy smoker,
- alcohol consumption,
- short/long hair,
- wearing higher heels or not,
- ..., etc.

iClicker Question 4.1

According to “Cumulative Use of Strong Anticholinergics and Incident Dementia” (JAMA, March 2015), from 10 years of tracking older adults and their use of anticholinergic drugs (meant to reduce symptoms of allergies, inability to sleep, anxiety, depression and bladder over-activity), the risk of Alzheimer’s was 63 percent higher. Which of the following is the **outcome variable** (response)?

1. use of anticholinergic drugs
2. risk of Alzheimer’s
3. hidden variables such as high blood pressure
4. none of the previous

iClicker Question 4.2

According to "Cumulative Use of Strong Anticholinergics and Incident Dementia" (JAMA, March 2015), from 10 years of tracking older adults and their use of anticholinergic drugs (meant to reduce symptoms of allergies, inability to sleep, anxiety, depression and bladder over-activity), the risk of Alzheimer's was 63 percent higher. Which of the following is the **probable cause**?

1. use of anticholinergic drugs
2. risk of Alzheimer's
3. hidden variables such as high blood pressure
4. none of the previous

Outline

Threats to Valid Comparisons

1. Hidden Confounder
2. In the Headlines

Examples In the Headlines

- Smoking ‘causes’ lung cancer. True? Having lung cancer or not is the outcome variable, smoking is probable cause, potential confounders abound. (see textbook)

Examples In the Headlines

- Smoking ‘causes’ lung cancer. True? Having lung cancer or not is the outcome variable, smoking is probable cause, potential confounders abound. (see textbook)
- not true for a single study or a few studies.

Examples In the Headlines

- "Older Viagra Users More Likely to Get STDs" – the Chicago Sun Times and the like. See critiques by Rebecca Goldin and Jing Peng in August 2010 from `http://www.stats.org` titled 'If you take Viagra, will you get an STD?' Pathway graph for this is 'person type' (i.e. lifestyle) is a confounder which is the cause (or surrogate of causes)

Examples In the Headlines

- "Older Viagra Users More Likely to Get STDs" – the Chicago Sun Times and the like. See critiques by Rebecca Goldin and Jing Peng in August 2010 from `http://www.stats.org` titled 'If you take Viagra, will you get an STD?' Pathway graph for this is 'person type' (i.e. lifestyle) is a confounder which is the cause (or surrogate of causes)
- 'taking ED drugs or not' is only a probable cause

Statistics and Data Analysis

STAT1600
Ch 5 Study Designs

Outline

Study Designs

- Randomized Trials
- Double-blind Randomized Controlled Trials
- Observational Studies
- iClicker Questions

Diet Comparison Example

The diet comparison example "Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women The A TO Z Weight Loss Study: A Randomized Trial" by C.D. Gardner, et al. (JAMA, Vol. 297, pp. 969–977, March 2007) is a good example of **randomized trials**.

An important characteristic of the experiment is that the comparison groups are similar to each other in all aspects, **except for the treatment** (see Table 5.1 on page 79). Hence such experiment offers fair comparison of the treatment among the four diet groups.

Randomization and Randomized Trials

- In a randomized trial, subjects entering the trial in a randomized fashion (using virtual roll of a die) into one of several treatment groups. This process is called:

Randomization and Randomized Trials

- In a randomized trial, subjects entering the trial in a randomized fashion (using virtual roll of a die) into one of several treatment groups. This process is called:
- Randomization

Randomization and Randomized Trials

- In a randomized trial, subjects entering the trial in a randomized fashion (using virtual roll of a die) into one of several treatment groups. This process is called:
- Randomization
- the best way to safeguard against potential confounders so that the comparison groups are similar in all factors except for the treatment itself.

Randomized Controlled Experiment (or Trial)

is a randomized experiment in which one of the comparison groups is a control group or placebo group.

Double-blind Randomized Controlled Trial

is a randomized controlled trial in which neither doctor (the experimenter) nor patient (experimental subject) knows what treatment the patient receives.

- done by giving the patient a pill that looks/smells... like the treatment pill, but is actually an inert pill or placebo.

Double-blind Randomized Controlled Trial

is a randomized controlled trial in which neither doctor (the experimenter) nor patient (experimental subject) knows what treatment the patient receives.

- done by giving the patient a pill that looks/smells... like the treatment pill, but is actually an inert pill or placebo.
- Blinding achieves additional protection against bias.

Double-blind Randomized Controlled Trial

is a randomized controlled trial in which neither doctor (the experimenter) nor patient (experimental subject) knows what treatment the patient receives.

- done by giving the patient a pill that looks/smells... like the treatment pill, but is actually an inert pill or placebo.
- Blinding achieves additional protection against bias.
- All groups have the same frame of mind (as opposed to knowing you are not really getting the new drug)

Double-blind Randomized Controlled Trial

is a randomized controlled trial in which neither doctor (the experimenter) nor patient (experimental subject) knows what treatment the patient receives.

- done by giving the patient a pill that looks/smells... like the treatment pill, but is actually an inert pill or placebo.
- Blinding achieves additional protection against bias.
- All groups have the same frame of mind (as opposed to knowing you are not really getting the new drug)
- The experimenter has the same frame of mind evaluating patients from each group.

Leg Fracture Example

In many cases, randomization cannot be achieved in that the treatments being compared cannot be assigned, e.g., the study involving women and leg fractures. When a new subject enters the study (by having a car accident), we observe what gender they belong to, instead of randomly assigning it. This is an example of *observational studies*.

Typical Reasons for Observational Studies

- Assigning treatment is impossible
  E.g. to compare fracture rates between men and women, we cannot randomize subjects into the comparison groups

Typical Reasons for Observational Studies

- Assigning treatment is impossible
  E.g. to compare fracture rates between men and women, we cannot randomize subjects into the comparison groups
- Assigning treatment is unethical
  E.g. to compare cancer rates of smokers and nonsmokers, we do not want to *randomize* subjects into smoker-nonsmoker comparison groups

Typical Reasons for Observational Studies

- Assigning treatment is impossible
  E.g. to compare fracture rates between men and women, we cannot randomize subjects into the comparison groups
- Assigning treatment is unethical
  E.g. to compare cancer rates of smokers and nonsmokers, we do not want to *randomize* subjects into smoker-nonsmoker comparison groups
- Assigning treatment is impractical
  E.g. the outcome is a rare event like cancer or stroke, and a randomized trial would need too many subjects and too much time.

Typical Reasons for Observational Studies

- Assigning treatment is impossible
  E.g. to compare fracture rates between men and women, we cannot randomize subjects into the comparison groups
- Assigning treatment is unethical
  E.g. to compare cancer rates of smokers and nonsmokers, we do not want to *randomize* subjects into smoker-nonsmoker comparison groups
- Assigning treatment is impractical
  E.g. the outcome is a rare event like cancer or stroke, and a randomized trial would need too many subjects and too much time.
- In cases like these, a case-control study is generally the way to go.

Case-control Study

starts with the outcome and then works backward to the type of treatment. For instance in the diet comparison trial, a case-control study would look for people in the population who lost weight, and then ask them what diet they used.

Case-control studies, compared to randomized controlled experiments,

- are frequently used because they are cheaper and easier to conduct;

Case-control Study

starts with the outcome and then works backward to the type of treatment. For instance in the diet comparison trial, a case-control study would look for people in the population who lost weight, and then ask them what diet they used.

Case-control studies, compared to randomized controlled experiments,

- are frequently used because they are cheaper and easier to conduct;
- are less time-consuming to conduct;

Case-control Study

starts with the outcome and then works backward to the type of treatment. For instance in the diet comparison trial, a case-control study would look for people in the population who lost weight, and then ask them what diet they used.

Case-control studies, compared to randomized controlled experiments,

- are frequently used because they are cheaper and easier to conduct;
- are less time-consuming to conduct;
- are able to conclude a link or 'association,' but are not able to prove 'causation;'

Case-control Study

starts with the outcome and then works backward to the type of treatment. For instance in the diet comparison trial, a case-control study would look for people in the population who lost weight, and then ask them what diet they used.

Case-control studies, compared to randomized controlled experiments,

- are frequently used because they are cheaper and easier to conduct;
- are less time-consuming to conduct;
- are able to conclude a link or 'association,' but are not able to prove 'causation;'
- provide initial evidence that can generate resources for more rigorous studies like double-blind randomized controlled trials.

Successful Story of a Case-control Study

The first study formally linking lung cancer to smoking was a 1950 case-control study "Smoking and Carcinoma of the Lung" by Richard Doll and A. Bradford Hill (British Medical Journal, 1950 September 30; 2(4682): page 739–748). This study led to numerous studies, and consequently, it is now accepted by the scientific community that smoking causes lung cancer.

Case-crossover Study

allows subjects in the treatment group ‘cross over’ to the control group an vice versa. That is, each subject can be their own control.

A successful example: a 1997 study linking cell phone use to car accidents: “Association between cellular-telephone calls and motor vehicle collisions” by D.A. Redelmeier and R.J. Tibshirani (The New England Journal of Medicine, 1997 Feb 13; vol 336, pp. 453–458).

iClicker Question 5.1

According to "Cumulative Use of Strong Anticholinergics and Incident Dementia" (JAMA, March 2015), from 10 years of tracking older adults and their use of anticholinergic drugs (meant to reduce symptoms of allergies, inability to sleep, anxiety, depression and bladder over-activity), the risk of Alzheimer's was 63 percent higher. What type of study is this?

1. randomized controlled experiment
2. case-control study
3. none of the previous

iClicker Question 5.2

Which of the following is *false* about a case-control study when it is compared to a randomized controlled trial?

1. case-control study is less time-consuming to conduct
2. case-control study is cheaper to conduct
3. case-control study can be used to determine causation
4. case-control study is easier to conduct

iClicker Question 5.3

A clinical trial was conducted in which 120 patients with similar clinical features were randomly divided into a control group and a treatment group, each consisting of 60 patients. What type of study this is?

1. randomized controlled trial
2. case-control study
3. none of the previous

iClicker Question 5.4

Western Michigan University offers a 1 credit course for freshmen, UNIV 1010, which teaches about university resources and study habits for success in college. It is an elective course and about half of the freshmen take it. WMU has studied the results by comparing the retention and GPAs of students who took this class against those who did not take this class. It was found that retention and GPAs were generally higher for those who took UNIV 1010. This evidence was put forth as proof that the course was successful and that it should be continued. What potential source(s) of bias have not been accounted for by WMU?

1. GPAs before college.
2. Lack of randomization in subject selection for UNIV 1010
3. Chosen majors of the students
4. All the above

Statistics and Data Analysis

STAT 1600
Ch. 6 The Normal Distribution

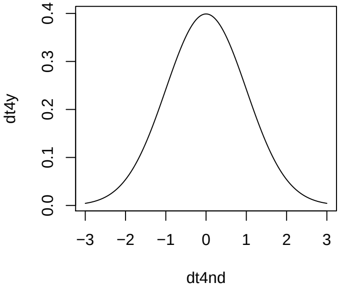
Outline

The Normal Distribution

- Normal Distribution and Z Score
- Using the Normal or Z Curve

Normal Distribution

- Normal distribution is denoted by N(mean,SD).
- Standard normal has mean=0 and SD=1 and is denoted by N(0, 1)z s
- The mean gives the location of the line of symmetry and the standard deviation refers to the spread.
- **The area under the curve always equals 1.**

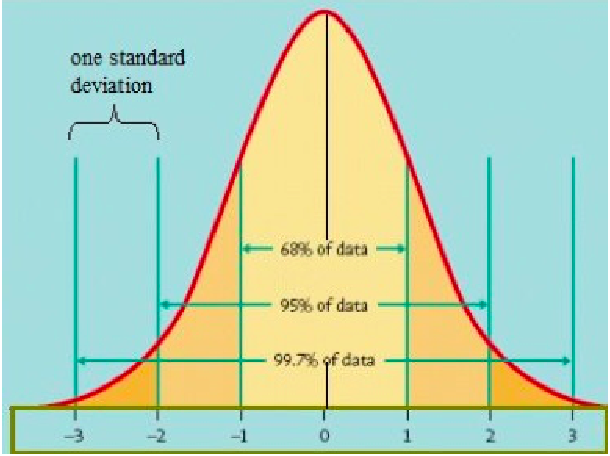


Normal Distribution

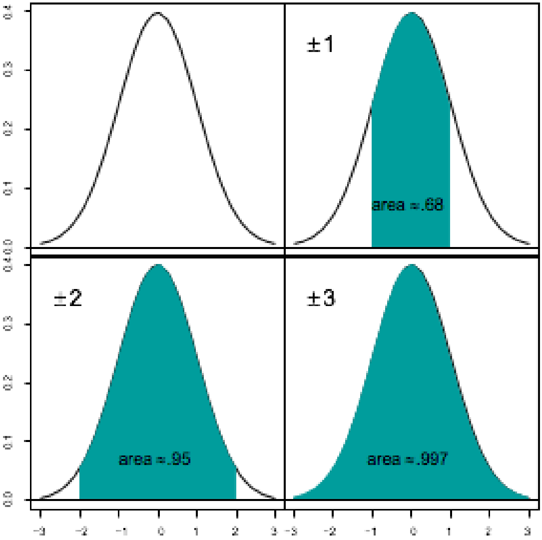
The normal or bell-shaped curve is helpful in calculating probabilities

- 68% of the data falls within -1 and +1 standard deviations of the mean
- 95% falls between -2 and +2 standard deviations
- 99.7% falls between -3 and +3 standard deviations

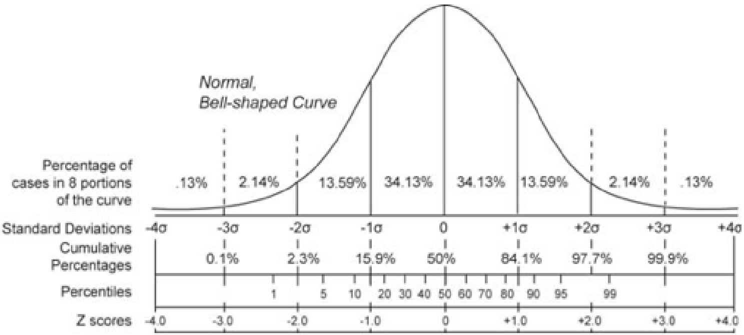
Normal Distribution



Normal Distribution



Normal Distribution



Z-score

- To calculate area under curve of general N(mean, SD), calculate z score (i.e., number of SD's above average/below the average). For example, the area to the left of X in this normal:

Z-score

- To calculate area under curve of general N(mean, SD), calculate z score (i.e., number of SD's above average/below the average). For example, the area to the left of X in this normal:
- $$\texttt{calculate z-score of x} = z = \frac{x - \mu}{\sigma}$$

Z-score

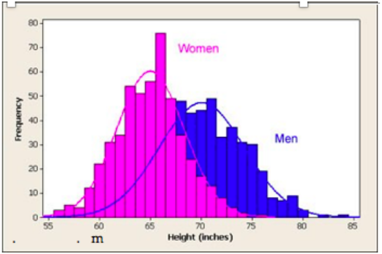
- To calculate area under curve of general N(mean, SD), calculate z score (i.e., number of SD's above average/below the average). For example, the area to the left of X in this normal:
- $$\texttt{calculate z-score of x} = z = \frac{x-\mu}{\sigma}$$
- To find the probability of a random variable, X, occurring in a normal distribution, we make use of the normal distribution or normal curve. Once we obtain a z-score using the formula above we can find the probability of a data value occuring at or below that value.

Z-score

- To calculate area under curve of general N(mean, SD), calculate z score (i.e., number of SD's above average/below the average). For example, the area to the left of X in this normal:
- $$\texttt{calculate z-score of x} = z = \frac{x - \mu}{\sigma}$$
- then area to the left of x in N(mean, SD) = area to the left of z in N(0,1)

Z-score

Let X =adult male height. Then X is N (70", 4"). This is stating that for this population the mean height in males is 70 inches and the standard deviation is 4 inches. What is the probability that a male is less than 6' tall (or 72 inches)?:

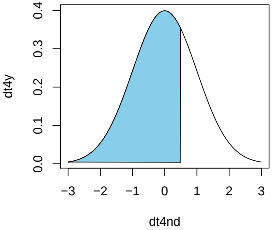


$$
\begin{aligned}
z &= \frac{x-\mu}{\sigma} \\
&= \frac{x-70}{4} \\
&= \frac{72-70}{4} \\
&= 0.5
\end{aligned}
$$

Z-Score

Let X = adult male height. Then X is N (70", 4"). This is stating that for this population the average height in males is 70 inches and the standard deviation is 4 inches. What is the probability that a male is less than 6' tall (or 72 inches)?:

$$\begin{aligned} z &= \frac{X-\mu}{\sigma} \\ &= \frac{X-70}{4} \\ &= \frac{72-70}{4} \\ &= 0.5 \end{aligned}$$



A Z value of 0.5 corresponds to the area of 0.6915 (AUC). This means the probability, $P(X \leq 72 \text{ inches})$, is 69.15%

Statistics and Data Analysis

STAT 1600
Ch 7 The Binomial Distribution

Outline

The Binomial Distribution

- Binomial Random Variables

Binomial Process and Binomial RV

A sequence of (fixed) n observations is called a **binomial process** if

Binomial Process and Binomial RV

A sequence of (fixed) n observations is called a **binomial process** if

- each observation results in exactly one of two possible outcomes (conveniently called *success* and *failure*)
- P (success) = p, and P (failure) = q = 1 - p for all observations
- observations are independent

Binomial Process and Binomial RV

A sequence of (fixed) n observations is called a **binomial process** if

- each observation results in exactly one of two possible outcomes (conveniently called *success* and *failure*)
- P (success) = p, and P (failure) = q = 1 - p for all observations
- observations are independent
- X = total number of successes among the n observations is a **binomial random variable** with parameters n and p and is denoted $X \sim binomial(n, p)$

Binomial Process and Binomial RV

Example: What is the probability of rolling exactly two ones in 10 rolls of a die? (n, our sample size = 10)
There are several things you need to know:

1. First Define Success: "Rolling a 1 on a single die"
2. Define the probability of success: (p):
   $p = \frac{1}{6} = 0.167$
3. Find the probability of failure:
   $(1 - p) = q = \frac{5}{6} = 0.833$
4. Define the X (or j) that we are investigating. This is the number of successes out of the trials(or sample size, n): Here it is two rolls out of 10 so $X = j = 2$

Binomial Process and Binomial RV

- Example: **What is the probability of rolling exactly two ones in 10 rolls of a die?** Anytime a '1' appears it is a success. Anytime any other number (2, 3, 4, 5, 6) appears it is a failure.

Binomial Process and Binomial RV

- Example: **What is the probability of rolling exactly two ones in 10 rolls of a die?** Anytime a '1' appears it is a success. Anytime any other number (2, 3, 4, 5, 6) appears it is a failure.
- We need to use the binomial probability distribution function in order to solve for this:

Binomial Process and Binomial RV

- Example: **What is the probability of rolling exactly two ones in 10 rolls of a die?** Anytime a '1' appears it is a success. Anytime any other number (2, 3, 4, 5, 6) appears it is a failure.
- We need to use the binomial probability distribution function in order to solve for this:
- $$P[X = j] = \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j},$$

  where j = 0, 1, 2, $\cdots$, n

Binomial Process and Binomial RV

In our example:
X = j = 2 and n = 10
p = (probability of success) = 1/6 = 0.167
q = (probability of failure) = 1 - p = 1 - 0.167 = 0.833
(Since q is the probability of failure in our example you can see it is also 5/6 as there are five other numbers the die could fall on.)

Using Formula to Compute Binomial Probability

$X \sim binomial(n, p)$

$$
\begin{aligned}
P[X = 2] &= \frac{10!}{2!(10-2)!}(.167)^2(.833)^{10-2} \\
&= \frac{10 \times 9 \times 8!}{(2 \times 1)(8!)}(.167)^2(.833)^8 \\
&= (45)(.167)^2(.833)^8 \\
&= 0.291 \texttt{ or } 29.1\%
\end{aligned}
$$

Examples of Binomial RV

- A 5-question multiple-choice quiz has 5 choices on each question. X = number of correct answers (success = correct) in the quiz by guessing all. Then $X \sim binomial(n = 5, p = 0.20)$.
- Past experience: 40% phone respondents agree to be interviewed (success = a respondent agrees to be interviewed) for market research survey. Of 50 reached by Reliable Research, X respondents agree to be interviewed. Then $X \sim binomial(n = 50, p = 0.40)$.

Examples of Binomial RV

- Suppose historical data shows that 20% of buyers at Best Buy who purchase smart fitness and GPS watches also purchase the Geek Squad's extended protection plan (success = a buyer purchases extended protection plan). X extended protection plans were sold along with the 300 smart watches sold last quarter. Then $X \sim binomial(n = 300, p = 0.20)$.

iClicker Question 7.1

The probability that a defective item is observed at a production line is 0.02. A quality engineer, working at the production line, inspects an item. What is the chance that the item is found to be non-defective?

1. 0.02
2. 1
3. 0.98
4. -0.02
5. none of the previous

iClicker Question 7.2

Over a long period of time in a large multinational corporation, 10% of all sales trainees are rated as outstanding, 75% are rated as excellent/good, 10% percent are rated as satisfactory, and 5% are considered unsatisfactory. What is the probability that a sales trainee is rated as not outstanding?

1. 0.05
2. 0.10
3. 0.25
4. 0.90
5. 0.95

Outline

The Binomial Distribution

- Binomial Random Variables
- Computing Binomial Probabilities Using a Formula

Using Formula to Compute Bin. Prob.

- $X \sim binomial(n, p)$
- $$P[X = j] = \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j}$$
- where j = 0, $\cdots$, n and $n! = n \times (n-1) \times \cdots \times 1$

Using Formula to Compute Bin. Prob.

- $X \sim binomial(n, p)$
- $$P[X = j] = \frac{n!}{j!(n-j)!} p^j (1-p)^{n-j}$$
- `where j = 0, ..., n` and $n! = n \times (n-1) \times \cdots \times 1$
- Multiple-choice quiz: $X \sim binomial(5, 0.2)$, eg.,
- $$\begin{aligned} P[X = 2] &= \frac{5!}{2!(5-2)!} 0.2^2 (1-0.2)^{5-2} \\ &= \frac{5 \cdot 4 \cdot 3!}{2 \cdot 1(3!)} 0.2^2 \cdot 0.8^3 = 0.2048 \end{aligned}$$

Best Buy Example, continued

Recall that historical data shows that 20% (i.e., $p = 0.2$) of buyers at Best Buy purchase extended protection plans with smart watches. If ($n = 10$ smart watches were sold in one day, what is the probability that ($j = 3$) extended protection plans were sold? Now, X , the number of extended protection plans sold along with 10 smart watches has $X \sim binomial(10, .2)$ distribution and hence

$$\begin{aligned} P[X = 3] &= \frac{10!}{3!(10-3)!} 0.2^3 (1-0.2)^{10-3} \\ &= \frac{10 \cdot 9 \cdot 8 \cdot 7!}{3 \cdot 2 \cdot 1(7!)} 0.2^3 \cdot 0.8^7 \\ &= 0.2013 \end{aligned}$$

Using Formula – olympics swimmer eg.,

A swimmer competes in three events in the Summer Olympics. The swimmer's winning/losing one event is independent of her result in any other event. If the probability of winning any one event is 0.45, what is the chance that she wins two or three events?
$X \sim binomial(3, 0.45)$

$$
\begin{aligned}
P[X = 2 or X = 3] &= P[X = 2] + P[X = 3] \\
&= \frac{3!}{2!(1)!} 0.45^2 0.55^1 + \frac{3!}{1!(0)!} 0.45^3 0.55^0 \\
&= 0.334125 + 0.091125 \\
&= 0.42525
\end{aligned}
$$

The ‘Language’ of Probability

- Note first that X , the number of successes, can only assume values 0, 1, ..., n.

The ‘Language’ of Probability

- Note first that X , the number of successes, can only assume values 0, 1, ..., n.
- ‘only 2’ or ‘exactly 2’: $P(X = 2)$

The 'Language' of Probability

- Note first that X , the number of successes, can only assume values 0, 1, ..., n.
- 'only 2' or 'exactly 2': P(X = 2)
- 'at most 3 or 'no more than 3' or '3 or less':
  $P[X \leq 3] = P(X = 0, 1, 2, \texttt{ or } 3) =$
  $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

The 'Language' of Probability

- Note first that X , the number of successes, can only assume values 0, 1, ..., n.
- 'only 2' or 'exactly 2': P(X = 2)
- 'at most 3 or 'no more than 3' or '3 or less': $P[X \leq 3] = P(X = 0, 1, 2, \texttt{ or } 3) =$ $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$
- 'at least 8' or 'no less than 8' or '8 or more' if n=10: $P[X \geq 8] =$ $P[X = 8] + P[X = 9] + P[X = 10]$

The 'Language' of Probability

- Note first that X , the number of successes, can only assume values 0, 1, ..., n.
- 'only 2' or 'exactly 2': P(X = 2)
- 'at most 3 or 'no more than 3' or '3 or less': $P[X \leq 3] = P(X = 0, 1, 2, \texttt{ or } 3) =$ $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$
- 'at least 8' or 'no less than 8' or '8 or more' if n=10: $P[X \geq 8] =$ $P[X = 8] + P[X = 9] + P[X = 10]$
- etc.

iClicker Question 7.3

The probability that a defective item is observed at a production line is 0.02. A quality engineer, working at the production line, goes to inspect the next 4 items. What is the set of possible number of defectives?

1. 1,2,3,4
2. 0,1,2,3,4
3. 1,2
4. 3,4
5. none of the previous

Statistics and Data Analysis

STAT1600
Ch. 8 Sampling Distribution of the Proportion

Outline

Distribution of the Sample Proportion

- Best Buy Example
- Theory
- Law of Large Numbers for Sample Proportions

Sampling Distribution of the Proportion

- Suppose Best Buy sells 60 extended protection plans with 300 smart watches sold.

Sampling Distribution of the Proportion

- Suppose Best Buy sells 60 extended protection plans with 300 smart watches sold.
- The protection plan sales rate is $\frac{60}{300} = 0.20$.

Sampling Distribution of the Proportion

- Suppose Best Buy sells 60 extended protection plans with 300 smart watches sold.
- The protection plan sales rate is $\frac{60}{300} = 0.20$.
- Therefore, let X denote the number of successes out of a sample of n observations. Then X is a binomial random variable with parameters n and p. Note that p is the (population) proportion of successes.

Sampling Distribution of the Proportion

- Suppose Best Buy sells 60 extended protection plans with 300 smart watches sold.
- The protection plan sales rate is $\frac{60}{300} = 0.20$.
- Therefore, let X denote the number of successes out of a sample of n observations. Then X is a binomial random variable with parameters n and p. Note that p is the (population) proportion of successes.
- The (sample) proportion of successes, $\hat{p} = \frac{x}{n}$ in a sample is also a random variable.

Sampling Distribution of the Proportion

- $\hat{p} = \frac{X}{n} =$ (number of successes) / (sample size)

Sampling Distribution of the Proportion

- $\hat{p} = \frac{X}{n} =$ (number of successes) / (sample size)
- For the binomial, X, the number of successes, is expected to be around np give or take $\sqrt{npq}$.

Sampling Distribution of the Proportion

- $\hat{p} = \frac{X}{n} =$ (number of successes) / (sample size)
- For the binomial, X, the number of successes, is expected to be around np give or take $\sqrt{npq}$.
- For the proportion, $\hat{p}$ is expected to be $p = \frac{n\hat{p}}{n}$ give or take $\sqrt{\frac{pq}{n}} = \frac{\sqrt{npq}}{n}$.

Sampling Distribution of the Proportion

- $\hat{p} = \frac{X}{n}$ = (number of successes) / (sample size)
- For the binomial, X, the number of successes, is expected to be around np give or take $\sqrt{npq}$.
- For the proportion, $\hat{p}$ is expected to be $p = \frac{n\hat{p}}{n}$ give or take $\sqrt{\frac{pq}{n}} = \frac{\sqrt{npq}}{n}$.
-

Random Variable	Mean	SD
X	np	$\sqrt{npq}$
$\hat{p}$	p	$\sqrt{\frac{pq}{n}}$

Best Buy example, revisited

- The number of protection plans sold is expected to be around 60 ± 7

Best Buy example, revisited

- The number of protection plans sold is expected to be around 60 ± 7
- The proportion of plans sold is expected to be around
$$\frac{60}{300} \pm \frac{7}{300} \texttt{ or } 0.2 \pm 0.02$$

Best Buy example, revisited

- The number of protection plans sold is expected to be around 60 ± 7
- The proportion of plans sold is expected to be around
$$\frac{60}{300} \pm \frac{7}{300} \quad \texttt{or} \quad 0.2 \pm 0.02$$
- The *percentage* of plans sold is expected to be around 20% give or take 2% (Note: percentage $=$ proportion $\times$ 100%)

Gamers Retro Rental Eg., revisited

- Historically, 5% of videogame rentals from Gamers Retro Rental are returned late.

Gamers Retro Rental Eg., revisited

- Historically, 5% of videogame rentals from Gamers Retro Rental are returned late.
- Gamers Retro Rental rented out 100 videogames yesterday. The percentage that will be returned late should be around 5%, give or take

$$100\% \times \sqrt{\frac{0.05 \times 0.95}{100}} \approx 2.2\%$$

Gamers Retro Rental Eg., revisited

- Historically, 5% of videogame rentals from Gamers Retro Rental are returned late.
- Gamers Retro Rental rented out 100 videogames yesterday. The percentage that will be returned late should be around 5%, give or take

$$100\% \times \sqrt{\frac{0.05 \times 0.95}{100}} \approx 2.2\%$$

- Gamers Retro Rental rented out 700 videogames yesterday. The percentage that will be returned late should be around 5%, give or take

$$100\% \times \sqrt{\frac{0.05 \times 0.95}{700}} \approx 0.8\%$$

iClicker Question 8.1

A study surveyed 100 students who took a standardized test. Among these students, 43 said they would like math help. What is the sample percentage of students needing math help?

1. 100%
2. 43%
3. 0.43%
4. 1%
5. cannot determine

Law of Large Numbers

- for sample proportions
- The sample proportion tends to get closer to the true proportion as sample size increases.

Law of Large Numbers

- for sample proportions
- The sample proportion tends to get closer to the true proportion as sample size increases.
- For the Best Buy Example:
- Recall if Best Buy sold 300 protection plans then $sd = 0.02$. Note that $p = 0.2$.

Law of Large Numbers

- for sample proportions
- The sample proportion tends to get closer to the true proportion as sample size increases.
- For the Best Buy Example:
- Recall if Best Buy sold 300 protection plans then $sd = 0.02$. Note that $p = 0.2$.
- If Best Buy sold 1200 plans then,

$$SD = \sqrt{\frac{0.2 \cdot 0.8}{1200}} = 0.0115$$

Sampling Distn of Sample Proportion

If Best Buy sold 100 protection plans with their smart watches last year, the percentage of watches sold with protection plans is expected to be around 20% give or take 4%. Estimate the likelihood that it sold a protection plan with each smart watch for more than 25% of those watches, in other words,

$$P[\hat{p} > 0.25] = ?$$

Sample Proportion is approx. normal

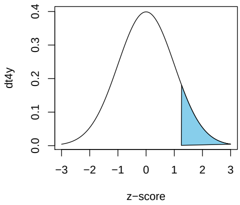
Given: $n = 100$ and $p = .2$

and $SD = \sqrt{\frac{.2(.8)}{100}} = 0.04$

and $P[\hat{p} > 0.25] = ?$

Note that $\hat{p} \approx N(0.2, 0.04)$

$z = \frac{0.25-0.2}{0.04} = 1.25$



$$
\begin{aligned}
P[\hat{p} > 0.25] &\approx P[Z > 1.25] \\
&\approx 1 - P[Z < 1.25] \\
&\approx 1 - .8944 \\
&\approx 0.1056
\end{aligned}
$$

iClicker Question 6.2

Recall that if Best Buy sold 100 smart watches last year, the percentage of watches sold with extended protection plans is expected to be around 20% give or take 4%. What is the chance that the percentage of plans sold with extended warranties is between $12\%(= 20\% - 2 \times 4\%)$ and $28\%(= 20\% + 2 \times 4\%)$?

1. 99.7%
2. 95%
3. 68%
4. 75%
5. cannot determine

Outline

Estimating Proportion

- Questions Asked About Population Proportion

Questions Asked

about the population proportion

- The population proportion p are generally unknown and are estimated from the data.

Questions Asked

about the population proportion

- The population proportion p are generally unknown and are estimated from the data.
- Suppose we want to estimate the number of students planning to attend graduate school.

Questions Asked

about the population proportion

- The population proportion p are generally unknown and are estimated from the data.
- Suppose we want to estimate the number of students planning to attend graduate school.
  1. Will the sample proportion equal the population proportion? Yes or No.

Questions Asked

about the population proportion

- The population proportion p are generally unknown and are estimated from the data.
- Suppose we want to estimate the number of students planning to attend graduate school.
    1. Will the sample proportion equal the population proportion? Yes or No.
    2. If not, by how much will it miss?

Estimating the population proportion p

- $\hat{p}$ is an estimate of the population proportion, i.e.,

$$E[\hat{p}] = p$$

Estimating the population proportion p

- $\hat{p}$ is an estimate of the population proportion, i.e.,

$$E[\hat{p}] = p$$

- Our estimate misses it by the standard error of the proportion

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Estimating the population proportion p

- $\hat{p}$ is an estimate of the population proportion, i.e.,

$$E[\hat{p}] = p$$

- Our estimate misses it by the standard error of the proportion

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Consider our example: $n = 40$ graduating seniors, $X = 6$ plan to attend graduate school.

Estimating the population proportion p

- $\hat{p}$ is an estimate of the population proportion, i.e.,

$$E[\hat{p}] = p$$

- Our estimate misses it by the standard error of the proportion

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Consider our example: $n = 40$ graduating seniors, $X = 6$ plan to attend graduate school.
  1. What is the proportion of graduating seniors planning to attend graduate school?

Estimating the population proportion p

- $\hat{p}$ is an estimate of the population proportion, i.e.,

$$E[\hat{p}] = p$$

- Our estimate misses it by the standard error of the proportion

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Consider our example: $n = 40$ graduating seniors, $X = 6$ plan to attend graduate school.
    1. What is the proportion of graduating seniors planning to attend graduate school?
    2. By how much will it miss the true population proportion?

Estimating the pop. proportion – Cont'd

- $\hat{p} = \frac{X}{n} = \frac{6}{40} = 0.15$

Estimating the pop. proportion – Cont'd

- $\hat{p} = \frac{X}{n} = \frac{6}{40} = 0.15$
- $$\begin{aligned} SE_{\hat{p}} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.15 \times .85}{40}} \\ &= 0.056 \end{aligned}$$

Estimating the pop. proportion – Cont'd

- $\hat{p} = \frac{X}{n} = \frac{6}{40} = 0.15$
- $$\begin{aligned} SE_{\hat{p}} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.15 \times .85}{40}} \\ &= 0.056 \end{aligned}$$
- What if 54 out of 360 students plan to go to graduate school. The proportion of all students who plan to go to graduate school is estimated as

Estimating the pop. proportion – Cont'd

- $\hat{p} = \frac{X}{n} = \frac{6}{40} = 0.15$
-
$$\begin{aligned} SE_{\hat{p}} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.15 \times .85}{40}} \\ &= 0.056 \end{aligned}$$

- What if 54 out of 360 students plan to go to graduate school. The proportion of all students who plan to go to graduate school is estimated as
- $\hat{p} = \frac{54}{360} = 0.15$ with $SE_{\hat{p}} = \sqrt{\frac{.15 \times .85}{360}} = .0188$

Estimating the pop. proportion – Cont'd

- The **population** proportion p is estimated using the sample proportion $\hat{p}$, i.e., $E[\hat{p}] = p$.

Estimating the pop. proportion – Cont'd

- The **population** proportion p is estimated using the sample proportion $\hat{p}$, i.e., $E[\hat{p}] = p$.
- This estimate tends to miss by an amount called the $SE_{\hat{p}}$.

Estimating the pop. proportion – Cont'd

- The **population** proportion p is estimated using the sample proportion $\hat{p}$, i.e., $E[\hat{p}] = p$.
- This estimate tends to miss by an amount called the $SE_{\hat{p}}$.
- The $SE_{\hat{p}}$ is calculated as

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Estimating the pop. proportion – Cont'd

- The **population** proportion p is estimated using the sample proportion $\hat{p}$, i.e., $E[\hat{p}] = p$.
- This estimate tends to miss by an amount called the $SE_{\hat{p}}$.
- The $SE_{\hat{p}}$ is calculated as

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- As sample size increases, the $SE_{\hat{p}}$ decreases.

iClicker Question 8.3

Which of the following statements is true about the standard error of the sample proportion?

1. The standard error increases when sample size increases.
2. The standard error decreases when sample size decreases.
3. The increase/decrease of sample size has no effect on the value of the standard error.
4. The standard error decreases when sample size increases.
5. None of the previous.

Statistics and Data Analysis

STAT 1600
Ch 9 Comparing Two Proportions,
Part 1 Difference in Proportions

Outline

Difference Between Independent Proportions

- Example and Notation
- Standard Error of Difference in Sample Proportions

Change in Student Retention Rate

- Has retention rate at WMU changing?

Change in Student Retention Rate

- Has retention rate at WMU changing?
- A random sample of 200 entering students in 1989 $\Rightarrow$ 74% were still enrolled 3 years later.

Change in Student Retention Rate

- Has retention rate at WMU changing?
- A random sample of 200 entering students in 1989 $\Rightarrow$ 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 $\Rightarrow$ 66% were still enrolled 3 years later.

Change in Student Retention Rate

- Has retention rate at WMU changing?
- A random sample of 200 entering students in 1989 $\Rightarrow$ 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 $\Rightarrow$ 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.

Change in Student Retention Rate

- Has retention rate at WMU changing?
- A random sample of 200 entering students in 1989 $\Rightarrow$ 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 $\Rightarrow$ 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.
- The 8% difference is based on random sampling, and is only an estimate of the true difference.

Change in Student Retention Rate

- Has retention rate at WMU changing?
- A random sample of 200 entering students in 1989 $\Rightarrow$ 74% were still enrolled 3 years later.
- Another random sample of 200 entering students in 1999 $\Rightarrow$ 66% were still enrolled 3 years later.
- An 8% change in 3-year retention rate was observed.
- The 8% difference is based on random sampling, and is only an estimate of the true difference.
- What is the likely size of the error of estimation?

Notation

A categorical variable with binary responses ('success' and 'failure') is of interest for two independent populations.

- Population 1 has proportion p_1 of successes.

Notation

A categorical variable with binary responses ('success' and 'failure') is of interest for two independent populations.

- Population 1 has proportion p_1 of successes.
- Population 2 has proportion p_2 of successes.

Notation

A categorical variable with binary responses ('success' and 'failure') is of interest for two independent populations.

- Population 1 has proportion p_1 of successes.
- Population 2 has proportion p_2 of successes.
- Sample of size n_1 is taken from population 1: X successes observed in the sample with sample proportion $\hat{p}_1 = \frac{X}{n_1}$

Notation

A categorical variable with binary responses ('success' and 'failure') is of interest for two independent populations.

- Population 1 has proportion p_1 of successes.
- Population 2 has proportion p_2 of successes.
- Sample of size n_1 is taken from population 1: X successes observed in the sample with sample proportion $\hat{p}_1 = \frac{X}{n_1}$
- Sample of size n_2 is taken from population 2: Y successes observed in the sample with sample proportion $\hat{p}_2 = \frac{Y}{n_2}$

Notation

A categorical variable with binary responses ('success' and 'failure') is of interest for two independent populations.

- Population 1 has proportion p_1 of successes.
- Population 2 has proportion p_2 of successes.
- Sample of size n_1 is taken from population 1: X successes observed in the sample with sample proportion $\hat{p}_1 = \frac{X}{n_1}$
- Sample of size n_2 is taken from population 2: Y successes observed in the sample with sample proportion $\hat{p}_2 = \frac{Y}{n_2}$
- The two samples are independent.

Standard Error of Difference

The SE (Standard Error) of the difference in the sample proportions of two independent samples is

$$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{(SE_{\hat{p}_1})^2 + (SE_{\hat{p}_2})^2}$$

where

$$SE_{\hat{p}_1} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}$$

$$SE_{\hat{p}_2} = \sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Change in Student Retention Rate

- For 1989 sample $\hat{p}_1 = 0.74$ give or take (i.e., with a standard error)

$$SE_{\hat{p}_1} = \sqrt{\frac{0.74(0.26)}{200}} = \sqrt{0.000962} = 0.031$$

Change in Student Retention Rate

- For 1989 sample $\hat{p}_1 = 0.74$ give or take (i.e., with a standard error)

$$SE_{\hat{p}_1} = \sqrt{\frac{0.74(0.26)}{200}} = \sqrt{0.000962} = 0.031$$

- For 1999 sample $\hat{p}_2 = 0.66$ give or take (i.e., with a standard error)

$$SE_{\hat{p}_2} = \sqrt{\frac{0.66(0.34)}{200}} = \sqrt{0.001122} = 0.033$$

Change in Student Retention Rate

- For 1989 sample $\hat{p}_1 = 0.74$ give or take (i.e., with a standard error)

$$SE_{\hat{p}_1} = \sqrt{\frac{0.74(0.26)}{200}} = \sqrt{0.000962} = 0.031$$

- For 1999 sample $\hat{p}_2 = 0.66$ give or take (i.e., with a standard error)

$$SE_{\hat{p}_2} = \sqrt{\frac{0.66(0.34)}{200}} = \sqrt{0.001122} = 0.033$$

- and hence for the difference in sample proportions.

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{0.000962 + 0.001122} = 0.0456$$

Calculation of the $SE_{\hat{p}_1-\hat{p}_2}$

- Calculate $(SE_1)^2$, the squared $SE_{\hat{p}_1}$

$$(SE_1)^2 = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}$$

keeping 6 decimal places to the right of the decimal point.

Calculation of the $SE_{\hat{p}_1 - \hat{p}_2}$

- Calculate $(SE_1)^2$, the squared $SE_{\hat{p}_1}$

$$(SE_1)^2 = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}$$

keeping 6 decimal places to the right of the decimal point.

- Calculate $(SE_2)^2$, the squared $SE_{\hat{p}_2}$

$$(SE_2)^2 = \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}$$

keeping 6 decimal places to the right of the decimal point.

Calculation of the $SE_{\hat{p}_1-\hat{p}_2}$

- Calculate $(SE_1)^2$, the squared $SE_{\hat{p}_1}$

$$(SE_1)^2 = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}$$

  keeping 6 decimal places to the right of the decimal point.

- Calculate $(SE_2)^2$, the squared $SE_{\hat{p}_2}$

$$(SE_2)^2 = \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$$

  keeping 6 decimal places to the right of the decimal point.

- Calculate $(SE_1)^2 + (SE_2)^2$.

Calculation of the $SE_{\hat{p}_1 - \hat{p}_2}$

- Calculate $(SE_1)^2$, the squared $SE_{\hat{p}_1}$

$$(SE_1)^2 = \frac{\hat{p}_1(1 - \hat{p}_1)}{n_1}$$

  keeping 6 decimal places to the right of the decimal point.

- Calculate $(SE_2)^2$, the squared $SE_{\hat{p}_2}$

$$(SE_2)^2 = \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}$$

  keeping 6 decimal places to the right of the decimal point.

- Calculate $(SE_1)^2 + (SE_2)^2$.
- $SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{(SE_1)^2 + (SE_2)^2}$

Calculation of the $SE_{\hat{p}_1-\hat{p}_2}$

Change in Retention Rate Example

$$
\begin{aligned}
(SE_1)^2 &= \frac{.74 \times .26}{200} = 0.000962 \\
(SE_2)^2 &= \frac{.66 \times .34}{200} = 0.001122 \\
(SE_1)^2 + (SE_2)^2 &= 0.000962 + 0.001122 = 0.002084 \\
SE_{\hat{p}_1-\hat{p}_2} &= \sqrt{0.002084} = 0.0456
\end{aligned}
$$

Outline

Confidence Interval for Difference in Proportions

- Confidence Interval for Difference in Proportions
- iClicker Questions

Confidence Interval for $\hat{p}_1 - \hat{p}_2$

A 95% confidence interval for the true difference $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \times SE_{\hat{p}_1 - \hat{p}_2}$$

That is

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

If the interval excludes zero (0), then we say that the difference in sample proportions is statistically significant.

However, If the interval includes 0 then the difference is statistically insignificant.

Change in Student Retention Rate

- Recall that the standard error of the difference in the sample proportions is
$$SE_{\hat{p}_1 - \hat{p}_2} = 0.0456$$

Change in Student Retention Rate

- Recall that the standard error of the difference in the sample proportions is
$$SE_{\hat{p}_1 - \hat{p}_2} = 0.0456$$
- So, a 95% CI (confidence interval) for $p_1 - p_2$ is
$(0.74 - 0.66) \pm 1.96 \times 0.0456 = 0.8 \pm 0.089 \Rightarrow$ (-.009, 0.169)

Change in Student Retention Rate

- Recall that the standard error of the difference in the sample proportions is
$$SE_{\hat{p}_1-\hat{p}_2} = 0.0456$$
- So, a 95% CI (confidence interval) for $p_1 - p_2$ is $(0.74 - 0.66) \pm 1.96 \times 0.0456 = 0.8 \pm 0.089 \Rightarrow$ (-.009, 0.169)
- If we round it off to (-.01, .17), or, in percentages, (-1%, 17%), we say that the drop in retention rate from 1989 to 1999 is between -1% and 17% with 95% confidence.

Change in Student Retention Rate

- Recall that the standard error of the difference in the sample proportions is
$$SE_{\hat{p}_1-\hat{p}_2} = 0.0456$$
- So, a 95% CI (confidence interval) for $p_1 - p_2$ is $(0.74 - 0.66) \pm 1.96 \times 0.0456 = 0.8 \pm 0.089 \Rightarrow$ (-.009, 0.169)
- If we round it off to (-.01, .17), or, in percentages, (-1%, 17%), we say that the drop in retention rate from 1989 to 1999 is between -1% and 17% with 95% confidence.
- Note: 0% is contained in this interval and hence there is still a probability that there might not be a real change in retention rate, just chance variability.

iClicker Question 9.1

A 95% confidence interval was constructed for the difference in the proportions $p_1 - p_2$ in two independent populations: (-0.08, 0.26). Which of the following is true?

1. The difference in the proportions is significant.
2. p_1 differs from p_2 significantly.
3. The difference in the proportions is insignificant.
4. None of the previous.

iClicker Question 9.2

A study of the television viewing preferences of children, each child is asked if the Sesame Street is the program he or she likes the best among others. Of 200 girls surveyed, 85 like Sesame Street the best; of 100 boys surveyed, 30 like Sesame Street the best. A 95% confidence interval for the difference in the percentages of children like the Sesame Street the best between girls and boys is (1.2%,23.8%).

1. Which of the following is true?
2. The two percentages differ significantly.
3. The two percentages do not differ significantly.
4. The two proportions do not differ significantly.
5. None of the previous.

Outline

Statistical Significance

Cooks or Chefs

- According to a 2009 occupation survey by the Census Bureau, regular cooks were a separate classification from chefs or head cooks:

Occupation	Women	Men	Total	% Women
Cooks	441	762	1203	37
Chefs	45	245	290	16

- The difference in percentage is approximately 21%.
- Is the difference in percentages just luck of the draw, or due to something else besides chance?

Cooks or Chefs – Cont'd

- If chance was at work, how likely we get a difference in proportions of 0.21?
- The chance of this occurs is small $\Rightarrow < 0.0001$. That is, less than 1 in 10,000. This chance of getting 0.21 by chance is called a P-value.
- But how do we know that this P-value is less than 0.0001?

Cooks or Chefs – Cont'd

- The SE for the difference in proportion is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{.37 \cdot .63}{1203} + \frac{.16 \cdot .84}{290}} = 0.026$$

- And hence the chance to get a difference beyond ± 0.078 ($=$ 3SE) is 0.003 ($=$ 1 - .997 by the empirical rule), or 3 in 1,000.
- Similarly, the chance to get a difference beyond ± 0.104 ($=$ 4SE) is $0.00006 < 0.0001$, or less than 1 in 10,000.
- Now, in our example, a difference of 0.21 is beyond 8 SE. This cannot be just chance variability. Something else is at work.
- Note: the probability of 0.00006 above was obtained by computer.

Statistical Significance, The P-Value

The general rule for P-value for the difference:

- If P-value $\leq .05$, the difference is **statistically significant**. (difference is at least 1.96SE in absolute value)
- If P-value $\leq .01$, the difference is called **highly significant**. (difference is at least 2.58SE in absolute value)
- If P-value $> .05$, the difference is **insignificant**. (difference is less than 1.96SE in absolute value)

iClicker Question 9.3

A 95% confidence interval was constructed for difference in the proportions $p_1 - p_2$ in two independent populations: (-0.04, 0.16). Which of the following is true?

1. The p-value ≤ 0.05, the difference is statistically significant.
2. The p-value ≤ 0.01, the difference is called highly significant.
3. The p-value > 0.05, the difference is insignificant.
4. None of the above.

Statistics and Data Analysis

Stat1600
Ch 9.3.2 Comparing Two Proportions,
Part 2 Risk Ratio

Outline

Risk Ratio

- Risk Ratio
- iClicker Questions

Hepatitis E Vaccine

From the study 'Safety and Efficacy of a Recombinant Hepatitis E Vaccine' by Shrestha et al. in the New England Journal of Medicine in March 2007 (Vol. 356 No. 9), the results were

	Hepatitis E		
	Yes	No	Total
Vaccine	3	895	898
Placebo	66	830	896

The vaccine efficacy, as reported in the article, was 95.5% with a 95% confidence interval of (85.6%, 98.6%).

How?

Risk ratio

Consider:

	Disease		
	Yes	No	Total
Exposure	a	b	a + b
No exposure	c	d	c + d

where Exposure = Exposure to treatment (i.e., Vaccine in this example).

The *risk ratio* (or relative risk) is

$$RR = \frac{P[Disease, exposure]}{P[Disease, Noexposure]} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

and the *efficacy* of the exposure is (1 - risk ratio) if `risk ratio` ≤ 1.

Hepatitis E Vaccine Eg., Cont'd

$$RR = \frac{\frac{3}{898}}{\frac{66}{896}} = 0.045 \texttt{ or } 4.5\%$$

That is, getting the vaccine reduces your risk to only 4.5% of the original. The efficacy of the vaccine is 95.5% (= 100% - 4.5%)

Calculating a 95% CI for RR

1. Calculate a 95% confidence interval for ln(RR):

Calculating a 95% CI for RR

1. Calculate a 95% confidence interval for ln(RR):
    1. calculate ln(RR)

Calculating a 95% CI for RR

1. Calculate a 95% confidence interval for ln(RR):
   1. calculate ln(RR)
   2. calculate

$$SE_{ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$

Calculating a 95% CI for RR

1. Calculate a 95% confidence interval for ln(RR):
   1. calculate ln(RR)
   2. calculate
   $$SE_{ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$
   3. calculate 95% CI for ln(RR)
   $$(ln(RR) - 1.96SE, ln(RR) + 1.96SE)$$

Calculating a 95% CI for RR

1. Calculate a 95% confidence interval for ln(RR):
   1. calculate ln(RR)
   2. calculate
   $$SE_{ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$
   3. calculate 95% CI for ln(RR)
   $$(ln(RR) - 1.96SE, ln(RR) + 1.96SE)$$

2. A 95% confidence interval for RR is
$$(e^{ln(RR)-1.96SE}, e^{ln(RR)+1.96SE})$$

Calculating a 95% CI for RR

1. Calculate a 95% confidence interval for ln(RR):
   1. calculate ln(RR)
   2. calculate
   $$SE_{ln(RR)} = \sqrt{\frac{1}{a} + \frac{1}{c} - \frac{1}{a+b} - \frac{1}{c+d}}$$
   3. calculate 95% CI for ln(RR)
   $$(ln(RR) - 1.96SE, ln(RR) + 1.96SE)$$
2. A 95% confidence interval for RR is
   $$(e^{ln(RR)-1.96SE}, e^{ln(RR)+1.96SE})$$
3. and RR is statistical significant if the interval excludes one (1).

Hepatits E Vaccine Example, Cont'd

1. Calculate a 95% confidence interval for ln(RR):

Hepatitis E Vaccine Example, Cont'd

1. Calculate a 95% confidence interval for ln(RR):
    1. calculate $\ln(RR) = \ln(.045) = -3.101$

Hepatits E Vaccine Example, Cont'd

1. Calculate a 95% confidence interval for ln(RR):
   1. calculate ln(RR) = ln(.045) = -3.101
   2. calculate

$$\begin{aligned} SE_{ln(RR)} &= \sqrt{\frac{1}{3}+\frac{1}{66}-\frac{1}{3+898}-\frac{1}{66+896}} \\ &= \sqrt{.3462} \\ &= 0.5884 \end{aligned}$$

Hepatits E Vaccine Example, Cont'd

1. Calculate a 95% confidence interval for ln(RR):
   1. calculate ln(RR) = ln(.045) = -3.101
   2. calculate

$$\begin{aligned} SE_{ln(RR)} &= \sqrt{\frac{1}{3} + \frac{1}{66} - \frac{1}{3+898} - \frac{1}{66+896}} \\ &= \sqrt{.3462} \\ &= 0.5884 \end{aligned}$$

   3. calculate 95% CI for ln(RR)

$$\begin{aligned} CI &= (-3.101 - 1.96 \cdot .5884, -3.101 + 1.96 \cdot .5884) \\ &= (-4.254, -1.948) \end{aligned}$$

Hepatits E Vaccine Example, Cont'd

1. A 95% confidence interval for RR is

$$(e^{-4.254}, e^{-1.948}) = (.014, .143)$$

Hepatits E Vaccine Example, Cont'd

1. A 95% confidence interval for RR is

$$(e^{-4.254}, e^{-1.948}) = (.014, .143)$$

2. That is, with 95% confidence, the relative risk of getting hepatitis with the vaccine is only 1.4% to 14.3% of placebo. In other words, the vaccine reduces your risk by as low as 85.7% (=100% - 14.3%) or as high as 98.6% (= 100% - 1.4%).

iClicker Question 9.2.1

In an observational study, a sample of 10000 smokers was taken, 50 were found to have lung cancer. Another sample of 10000 non-smokers was taken, only 2 have lung cancer. What is the relative risk of having lung cancer for smokers versus non-smokers?

1. 50
2. 2
3. 25
4. 100
5. cannot determine

iClicker Question 9.2.2

In an observational study, a sample of 10000 smokers was taken, 50 were found to have lung cancer. Another sample of 10000 non-smokers was taken, only 2 have lung cancer. A 95% confidence interval for the relative risk of having lung cancer for smokers versus smokers is (6.1, 102.5). Which of the following is true?

1. The proportion of smokers having lung cancer is significantly different than that of non-smokers.
2. There is no difference between the two proportions.
3. cannot determine

Statistics and Data Analysis

Stat 1600

Ch 9.3.3 Comparing Two Proportions, Part 3 Odds Ratio

Outline

- Odds Ratio
- iClicker Questions

Odds

The odds that an event occurs is

$$Odds = \frac{\texttt{Probability that an event occurs}}{\texttt{Probability that event does not occur}} = \frac{p}{1-p}$$

Probability	Odds
0.10	$\frac{1}{9} = 0.11$
0.20	$\frac{1}{4} = 0.25$
0.50	$\frac{1}{1} = 1.00$
0.80	$\frac{4}{1} = 4.00$
0.90	$\frac{9}{1} = 9.00$

iClicker Question 9.1

A clinical trial was conducted to study the efficacy of a new drug intended to lower the LDL (low-density lipoprotein, a.k.a., bad cholesterol). All study subjects showed a high LDL level at the baseline. Of the 100 treated subjects, 80 of them showed reduction to normal LDL level two weeks after treatment. What is the odds that a treated subject was having a reduction in LDL to normal level after two weeks?

1. 0.25
2. 4
3. 80
4. 20
5. cannot determine

Outline

- Odds Ratio
- iClicker Questions

Odds Ratio

When comparing two groups, the odds ratio is

$$OR = \frac{\texttt{Odds of group 1}}{\texttt{Odds of group 2}}$$

It is easier to interpret an OR when it's greater than 1 and hence, when $OR < 1$, exchange the roles of the groups above to get an $OR > 1$.

Hepatitis E Vaccine Example, revisited

	Hepatitis E		
	Yes	No	Total
Placebo	66	830	896
Vaccine	3	895	898

If we consider the placebo (i.e., unvaccinated) group as group 1, then

$$Odds(Hep|Placebo) = \frac{\frac{66}{896}}{\frac{830}{896}} = \frac{66}{830} = .07952$$

$$Odds(Hep|vaccine) = \frac{\frac{3}{898}}{\frac{895}{898}} = \frac{3}{895} = .00335$$

$$OddsRatio = \frac{\texttt{unvaccinated}}{\texttt{vaccinated}} = \frac{.07952}{.00335} = 23.7$$

So, the odds of getting hepatitis is about 24 times greater if you remain unvaccinated.

Hepatits E Vaccine Example, Cont'd

Consider,

	Disease		
	Yes	No	Total
Group 1	a	b	a + b
Group 2	c	d	c + d

The odds ratio is then

$$OR = \frac{a \times d}{b \times c} = \frac{success_1 \times failure_2}{failure_1 \times success_2} = \frac{success_1 / failure_1}{success_2 / failure_2}$$

where $successes_1$ = # of 'successes' (yes's) in group 1 and $failures_1$ = # of 'failures' (no's) in group 1;
$successes_2$ = # of 'successes' (yes's) in group 2 and $failures_2$ = # of 'failures' (no's) in group 2.

Note: assume $OR > 1$. If $OR < 1$ then switch the rows above and then the new OR is the reciprocal of the old one.

iClicker Question 9.2

A clinical trial was conducted to study the efficacy of a new drug intended to lower the LDL (low-density lipoprotein, a.k.a., bad cholesterol). All study subjects showed high LDL level at the baseline. Of the 100 treated subjects, 80 of them showed reduction to normal LDL level two weeks after treatment. On the other hand, of the 100 subjects who received placebo, only 10 showed reduction to normal LDL level after two weeks. What is the odds ratio of the treatment group versus the placebo group?

1. 36
2. 4
3. 9
4. 8
5. cannot determine

A 95% Confidence Interval for Odds Ratio

1. Calculate a 95% confidence interval for the ln odds ratio

A 95% Confidence Interval for Odds Ratio

1. Calculate a 95% confidence interval for the ln odds ratio
   - calculate ln odds ratio

$$ln(OR) = \frac{ad}{bc}$$

A 95% Confidence Interval for Odds Ratio

1. Calculate a 95% confidence interval for the ln odds ratio
   - calculate ln odds ratio

$$ln(OR) = \frac{ad}{bc}$$

   - calculate the standard error of ln(odds ratio)

$$SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

A 95% Confidence Interval for Odds Ratio

1. Calculate a 95% confidence interval for the ln odds ratio
   - calculate ln odds ratio

$$ln(OR) = \frac{ad}{bc}$$

   - calculate the standard error of ln(odds ratio)

$$SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

   - a 95% CI for ln(OR):

$$(ln(OR) - 1.96(SE), ln(OR) + 1.96(SE))$$

A 95% Confidence Interval for Odds Ratio

1. Calculate a 95% confidence interval for the ln odds ratio
   - calculate ln odds ratio
   $$ln(OR) = \frac{ad}{bc}$$
   - calculate the standard error of ln(odds ratio)
   $$SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$
   - a 95% CI for ln(OR):
   $$(ln(OR) - 1.96(SE), ln(OR) + 1.96(SE))$$

2. A 95% confidence interval for OR is
$$(e^{ln(OR)-1.96(SE)}, e^{ln(OR)+1.96(SE)})$$

Hepatits E Vaccine Example, revisited

- Recall that the odds ratio of placebo subjects getting hepatitis against that of vaccinated subjects is

Hepatits E Vaccine Example, revisited

- Recall that the odds ratio of placebo subjects getting hepatitis against that of vaccinated subjects is
- calculate ln odds ratio

$$ln(OR) = \frac{66 \times 895}{830 \times 3} = 23.7$$

Hepatits E Vaccine Example, revisited

- Recall that the odds ratio of placebo subjects getting hepatitis against that of vaccinated subjects is
- calculate ln odds ratio

$$ln(OR) = \frac{66 \times 895}{830 \times 3} = 23.7$$

- 95% CI for ln(OR):

Hepatits E Vaccine Example, revisited

- Recall that the odds ratio of placebo subjects getting hepatitis against that of vaccinated subjects is
- calculate ln odds ratio

$$ln(OR) = \frac{66 \times 895}{830 \times 3} = 23.7$$

- 95% CI for ln(OR):
-

$$ln(OR) = ln(23.7) = 3.165$$

Hepatits E Vaccine Example, revisited

- Recall that the odds ratio of placebo subjects getting hepatitis against that of vaccinated subjects is
- calculate ln odds ratio

$$ln(OR) = \frac{66 \times 895}{830 \times 3} = 23.7$$

- 95% CI for ln(OR):
-

$$ln(OR) = ln(23.7) = 3.165$$

- calculate the standard error of ln(odds ratio)

$$SE = \sqrt{\frac{1}{66} + \frac{1}{830} + \frac{1}{3} + \frac{1}{895}} = .5923$$

Hepatits E Vaccine Example, revisited

- a 95% CI for ln(OR):

$$(3.165 - 1.96(.5923), 3.165 + 1.96(.5923)) \Rightarrow (2.004, 4.326)$$

Hepatits E Vaccine Example, revisited

- a 95% CI for ln(OR):

$$(3.165 - 1.96(.5923), 3.165 + 1.96(.5923)) \Rightarrow (2.004, 4.326)$$

- A 95% confidence interval for OR is

$$(e^{2.004}, e^{4.326}) \Rightarrow (7.4, 75.6)$$

Hepatits E Vaccine Example, revisited

- a 95% CI for ln(OR):

$$(3.165 - 1.96(.5923), 3.165 + 1.96(.5923)) \Rightarrow (2.004, 4.326)$$

- A 95% confidence interval for OR is

$$(e^{2.004}, e^{4.326}) \Rightarrow (7.4, 75.6)$$

- So, with 95% confidence, the odds of *unvaccinated* subjects getting hepatitis is approximately between 7 and 76 times greater than that of the vaccinated subjects.

Statistics and Data Analysis

Stat 1600
Ch. 10 Sampling Distribution of the Mean

Outline

Sampling Distribution of the Mean

- Sample Mean Versus Individual Values
- Sampling From Normal Population
- iClicker Questions

Intro to the Sampling Distr'n of the Mean

- If we look at variability in the x variable (individual values) we would notice some extreme values

Intro to the Sampling Distr'n of the Mean

- If we look at variability in the x variable (individual values) we would notice some extreme values
- If we take a random sample of size n from the population we can calculate the sample mean, $\bar{X}$.

Intro to the Sampling Distr'n of the Mean

- If we look at variability in the x variable (individual values) we would notice some extreme values
- If we take a random sample of size n from the population we can calculate the sample mean, $\bar{X}$.
- The mean for our sample has to account for extremes in the data but if we take many more samples of the same size we would see the variability in the possible sample means will be less than the variability for the individual X values.

Intro to the Sampling Distr'n of the Mean

- This is because any extreme value will be averaged with other values in the sample.

Intro to the Sampling Distr'n of the Mean

- This is because any extreme value will be averaged with other values in the sample.
- Therefore, as you increase the size of the sample, you have more information; consequently, the sample mean is more accurate.

Example

Start with a population of numbers.
[1, 2, 3, 4, 5]

$$\texttt{Mean}\ \ \mu = 3, \texttt{SD}\ \ \sigma = 1.41$$

Consider the process of taking a random sample of size $n = 3$ from the box and calculating the sample mean $\bar{x}$.

1. Sample 1: [2, 4, 5]; $\bar{x}_1 = 3.67$

Example

Start with a population of numbers.
[1, 2, 3, 4, 5]

$$\texttt{Mean}\ \ \mu = 3, \texttt{SD}\ \ \sigma = 1.41$$

Consider the process of taking a random sample of size $n = 3$ from the box and calculating the sample mean $\bar{x}$.

1. Sample 1: [2, 4, 5]; $\bar{x}_1 = 3.67$
2. Sample 2: [1, 3, 5]; $\bar{x}_2 = 3.00$

Example

Start with a population of numbers.
[1, 2, 3, 4, 5]

$$\texttt{Mean}\ \ \mu = 3, \texttt{SD}\ \ \sigma = 1.41$$

Consider the process of taking a random sample of size $n = 3$ from the box and calculating the sample mean $\bar{x}$.

1. Sample 1: [2, 4, 5]; $\bar{x}_1 = 3.67$
2. Sample 2: [1, 3, 5]; $\bar{x}_2 = 3.00$
3. Sample 3: [1, 3, 4]; $\bar{x}_3 = 2.67$

Example

Start with a population of numbers.
[1, 2, 3, 4, 5] Mean $\mu = 3$, SD $\sigma = 1.41$
Consider the process of taking a random sample of size $n = 3$ from the box and calculating the sample mean $\bar{x}$.

- Sample 1: [2, 4, 5]; $\bar{x}_1 = 3.67$
- Sample 2: [1, 3, 5]; $\bar{x}_2 = 3.00$
- Sample 3: [1, 3, 4]; $\bar{x}_3 = 2.67$
- Sample 4: [1, 4, 5]; $\bar{x}_4 = 3.33$
- Sample 5: [2, 3, 4]; $\bar{x}_5 = 3.00$

Notice the different values for $\bar{X}_i$!

The value we would get for $\bar{x}$ is random, depending on chance.

Different samples yield different values for $\bar{x}$.

Standard Error of the Mean

Standard Error of the sample mean is the variation in $\bar{X}$:

$$\sigma_{\bar{x}} = SE_{\bar{x}} = \frac{SD}{\sqrt{n}}$$

where SD is the variability (the standard deviation) in the individual values and n is the sample size.

Note: The sample mean $\bar{X}$ has expected value of μ (the population mean). That is, the average of the sample means of all size-n samples is the population mean.

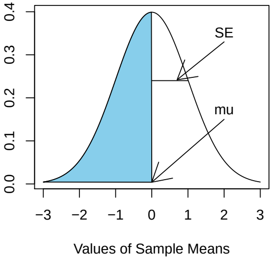
Sampling Distribution of a Sample Mean

The expected value of $\bar{x}$ is

$$E[\bar{x}] = \mu$$

The standard error of the mean is

$$SE = \frac{\sigma}{\sqrt{n}}$$



The sampling distribution is approximately normal (recommend $n > 30$).

Men's Height Example

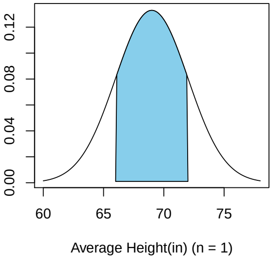
selecting one man

Suppose that male's height is approximately:

$N(mean = 69, SD = 3)$.

Using empirical rule

$0.68 \approx P(66 \leq X \leq 72)$

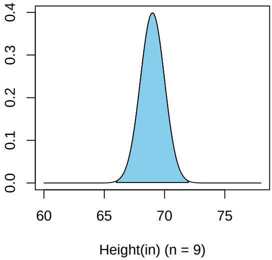


Men's Height Example

selecting nine men

Nine ($n = 9$) men are randomly selected. Now, $SE = \frac{3}{\sqrt{9}}$. Using empirical rule,

$0.997 \approx P[66 \leq X \leq 72]$



Men's Height Example; discussion

Therefore, when considering sample sizes of 1 or 9, our probability went from 68.27% to 99.73%.

The reason for this difference is that it is harder to get the mean height of 9 men to be less than 66 or greater than 72 versus for a single male

Distribution of the Sample Mean

when sampling from normal population.
If a population has a normal shaped histogram with $mean = \mu$, and standard deviation, $SD = \sigma$, then n-member averages will have a normal shaped histogram with $mean = \mu$ and $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
If $X_1, X_2, \cdots, X_n$ are sampled from $N(\mu, \sigma)$ then

$$X \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

.

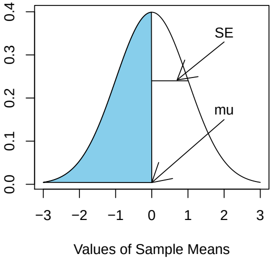
Sampling Distribution of a Sample Mean

The expected value of $\bar{x}$ is

$$E[\bar{x}] = \mu$$

The standard error of the mean is

$$SE = \frac{\sigma}{\sqrt{n}}$$



The sampling distribution is approximately normal (recommend $n > 30$).

Men's Height Example 1

a. $n = 1$; $P[X > 71] =?$; $z = \frac{71-69}{3} = 0.67$

   $P[X > 71] = P[z > 0.67] = 1 - P[z \leq 0.67] = 1 - .7486 = .2514$

Men's Height Example 1

a. $n = 1;\ P[X > 71] =?;\ z = \frac{71-69}{3} = 0.67$

   $P[X > 71] = P[z > 0.67] = 1 - P[z \leq 0.67] = 1 - .7486 = .2514$

b. $n = 1;\ P[X > 71] = 0.2514$

Men's Height Example 1

a. $n = 1$; $P[X > 71] =?$; $z = \frac{71-69}{3} = 0.67$
   $P[X > 71] = P[z > 0.67] = 1 - P[z \leq 0.67] = 1 - .7486 = .2514$
b. $n = 1$; $P[X > 71] = 0.2514$
c. $n = 9$; $P[\bar{X} > 71] =?$; $SE = \frac{3}{\sqrt{9}} = 1$; $z = \frac{71-69}{1} = 2$
   $P[X > 71] = P[z > 2] = 1 - P[z \leq 2] = 1 - .9772 = .0228$

Men's Height Example 1

a. $n = 1$; $P[X > 71] =?$; $z = \frac{71-69}{3} = 0.67$
   $P[X > 71] = P[z > 0.67] = 1 - P[z \leq 0.67] = 1 - .7486 = .2514$
b. $n = 1$; $P[X > 71] = 0.2514$
c. $n = 9$; $P[\bar{X} > 71] =?$; $SE = \frac{3}{\sqrt{9}} = 1$; $z = \frac{71-69}{1} = 2$
   $P[X > 71] = P[z > 2] = 1 - P[z \leq 2] = 1 - .9772 = .0228$
d. $n = 9$; $P[\bar{X} > 71] = 0.0228$

Men's Height Example 1

a. $n = 1$; $P[X > 71] = ?$; $z = \frac{71-69}{3} = 0.67$
   $P[X > 71] = P[z > 0.67] = 1 - P[z \leq 0.67] = 1 - .7486 = .2514$
b. $n = 1$; $P[X > 71] = 0.2514$
c. $n = 9$; $P[\bar{X} > 71] = ?$; $SE = \frac{3}{\sqrt{9}} = 1$; $z = \frac{71-69}{1} = 2$
   $P[X > 71] = P[z > 2] = 1 - P[z \leq 2] = 1 - .9772 = .0228$
d. $n = 9$; $P[\bar{X} > 71] = 0.0228$
e. $n = 25$; $P[\bar{X} > 71] = ?$; $SE = \frac{3}{\sqrt{25}} = .6$;
   $z = \frac{71-69}{.6} = 3.33$
   $P[\bar{X} > 71] = P[z > 3.33] = 1 - P[z \leq 3.33] = 1 - .9996 = .0004$

Men's Height Example 1

a. $n = 1;\ P[X > 71] = ?;\ z = \frac{71-69}{3} = 0.67$
   $P[X > 71] = P[z > 0.67] = 1 - P[z \le 0.67] = 1 - .7486 = .2514$
b. $n = 1;\ P[X > 71] = 0.2514$
c. $n = 9;\ P[\bar{X} > 71] = ?;\ SE = \frac{3}{\sqrt{9}} = 1;\ z = \frac{71-69}{1} = 2$
   $P[X > 71] = P[z > 2] = 1 - P[z \le 2] = 1 - .9772 = .0228$
d. $n = 9;\ P[\bar{X} > 71] = 0.0228$
e. $n = 25;\ P[\bar{X} > 71] = ?;\ SE = \frac{3}{\sqrt{25}} = .6;$
   $z = \frac{71-69}{.6} = 3.33$
   $P[\bar{X} > 71] = P[z > 3.33] = 1 - P[z \le 3.33] = 1 - .9996 = .0004$
f. $n = 25;\ P[\bar{X} > 71] = 0.0004$

Men's Height Example 1

a. $n = 9$; $P[\bar{X} > a] = .9$; $a = ?$
   Note that $P[z \geq -1.28] = 0.8997 \approx 0.9$.
   $P[\bar{X} > a] = P\left[z > \frac{a-69}{\sqrt{\frac{3}{\sqrt{9}}}}\right] \approx .90$
   $z \approx -1.28$
   $-1.28 = \frac{a-69}{1} \Rightarrow a = 69 + 1(-1.28) = 67.72$

Men's Height Example 1

a. $n = 9$; $P[\bar{X} > a] = .9$; $a = ?$
   Note that $P[z \geq -1.28] = 0.8997 \approx 0.9$.
   $$P[\bar{X} > a] = P\left[z > \frac{a-69}{\sqrt{\frac{3}{\sqrt{9}}}}\right] \approx .90$$
   $z \approx -1.28$
   $-1.28 = \frac{a-69}{1} \Rightarrow a = 69 + 1(-1.28) = 67.72$
b. $P[\bar{X} > a] = 0.9 \Rightarrow a = 67.72$

iClicker Question 10.1

It is suggested that the substrate concentration (mg/cm^3) of influent to a domestic-waste biofilm reactor is normally distributed with mean 0.30 and standard deviation of 0.06. A sample of 9 reactors is taken. What is the chance that the mean substrate concentration of influent of these reactors is in between $0.26 mg/cm^3$ and $0.34 mg/cm^3$?

1. 68%
2. 90%
3. 95%
4. 99.7%
5. none of the previous

Outline

Estimating the Population Mean μ

- Estimating the Population Mean μ
- iClicker Questions

Estimating the Population Mean

- The population mean μ is estimated using the sample mean $\bar{X}$.
- This estimate tends to miss by an amount called the standard error ($SE_{\bar{x}}$) of the mean.
- This is calculated as $\frac{SD}{\sqrt{n}}$. Note that SD is the sample standard deviation
- $$SD = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$$

WMU Undergraduates' Average GPA

- Suppose a sample of $n = 25$ WMU students yielded an average GPA of $\bar{X} = 3.05$ and a standard deviation of 0.40. Then the WMU true population average GPA, μ, is estimated by 3.05 with a standard error of $\frac{SD}{\sqrt{n}}$. What is the SE?

WMU Undergraduates' Average GPA

- Suppose a sample of $n = 25$ WMU students yielded an average GPA of $\bar{X} = 3.05$ and a standard deviation of 0.40. Then the WMU true population average GPA, μ, is estimated by 3.05 with a standard error of $\frac{SD}{\sqrt{n}}$. What is the SE?
- $\frac{SD}{\sqrt{n}} = \frac{0.40}{\sqrt{25}} = 0.08$

WMU Undergraduates' Average Stay

- A sample of $n = 25$ graduating students were randomly selected and asked about their length of stay.

WMU Undergraduates' Average Stay

- A sample of $n = 25$ graduating students were randomly selected and asked about their length of stay.
- Suppose that the sample averaged 5.3 years, with an SD of 1.5 years.

WMU Undergraduates' Average Stay

- A sample of $n = 25$ graduating students were randomly selected and asked about their length of stay.
- Suppose that the sample averaged 5.3 years, with an SD of 1.5 years.
- Then the true WMU average stay, μ, is estimated as $\bar{X} = 5.3$ years give or take $SE = \frac{SD}{\sqrt{n}} = \frac{1.5}{\sqrt{25}} = 0.3$ years.

WMU Undergraduates' Average Stay

- A second sample of 100 students were interviewed.

WMU Undergraduates' Average Stay

- A second sample of 100 students were interviewed.
- The mean and SD for the second sample were also 5.3 years and 1.5 years, respectively.

WMU Undergraduates' Average Stay

- A second sample of 100 students were interviewed.
- The mean and SD for the second sample were also 5.3 years and 1.5 years, respectively.
- Then the true average stay μ is estimated as $\bar{X} = 5.3$ years give or take $SE = \frac{SD}{\sqrt{n}} = \frac{1.5}{\sqrt{100}} = 0.15$ years.

WMU Undergraduates' Average Stay

- A second sample of 100 students were interviewed.
- The mean and SD for the second sample were also 5.3 years and 1.5 years, respectively.
- Then the true average stay μ is estimated as $\bar{X} = 5.3$ years give or take $SE = \frac{SD}{\sqrt{n}} = \frac{1.5}{\sqrt{100}} = 0.15$ years.
- Effect of Sample Size on Standard Error

WMU Undergraduates' Average Stay

- A second sample of 100 students were interviewed.
- The mean and SD for the second sample were also 5.3 years and 1.5 years, respectively.
- Then the true average stay μ is estimated as $\bar{X} = 5.3$ years give or take $SE = \frac{SD}{\sqrt{n}} = \frac{1.5}{\sqrt{100}} = 0.15$ years.
- Effect of Sample Size on Standard Error
- The standard error of the mean decreases like the square root of the sample size.

iClicker Question 10.2

The HDL (high-density lipoprotein, a.k.a., good cholesterol) among adults is normally distributed. A sample of 25 adults was selected which yielded a sample standard deviation of $5.3 mg/dL$. A second sample of 100 adults was selected which also yielded a sample standard deviation of $5.3 mg/dL$. Which of the following is true about the SEs (standard errors of the mean) of the two samples? $(SE_1 = SE_{n=25}, SE_2 = SE_{n=100})$

1. $SE_1 = 4SE_2$
2. $SE_2 = 2SE_1$
3. $SE_1 = 2SE_2$
4. $SE_2 = 4SE_1$
5. none of the previous

Outline

Sampling Distribution of the Mean

- Confidence Interval for Population Mean

Confidence Interval for Pop. Mean

A 95% confidence interval for μ

$$\bar{X} \pm 1.96 \frac{SD}{\sqrt{n}}$$

1. That is, first calculate the margin of error:

$$ME = 1.96 \times SE = 1.96 \frac{SD}{\sqrt{n}}$$

2. Then a 95% confidence interval is

$$(\bar{X} - ME, \bar{X} + ME)$$

WMU Undergraduates' Average GPA

Recall that the sample of $n = 25$ students yielded a sample mean of $\bar{X} = 3.05$ with a standard error of $SE = 0.08$. The margin of error is then $ME = 1.96 \times 0.08 = 0.157$. Hence a 95% CI for the true average GPA, μ, is:

$$(3.05 - 0.157, 3.05 + 0.157) = (2.893, 3.207)$$

.

Interpretation: based on the sample, we are 95% confident that the true mean GPA is in between 2.9 and 3.2, approximately.

Statistics and Data Analysis

Stat 1600
Ch. 11 Comparing Two Means

Outline

Comparing Two Means

- Comparing Means of Two Independent Populations
- Estimating the Difference
- iClicker Questions
- The P-Value

Two-Sample Problem

Two *independent* populations are to be compared for a characteristic, either a quantitative meansurement or a categorical one. A random sample is taken from each population to study. The samples are independent. In a controlled study, the subjects are selected with uniformity with respect to known source(s) of variation (similar weight, similar health condition, etc).

Examples

- Is there "grade inflation" in WMU? How does the average GPA of WMU students today compare with 10 years ago? Suppose a random sample of 100 student records from 10 years ago yields a sample average GPA of 2.90 with a standard deviation of 0.40. A random sample of 100 current students today yields a sample average of 2.98 with a standard deviation of .45. The difference between the two sample means is $2.98 - 2.90 = .08$. Is this proof that GPA's are higher today than 10 years ago? Or the chance variability is at work?
- How does the reduction of BMI of the Atkins diet compare to that of Zone diet at 12 months?

Notation

- Population 1 has mean μ_1 and standard deviation σ_1 (usually unknown).

Notation

- Population 1 has mean μ_1 and standard deviation σ_1 (usually unknown).
- Population 2 has mean μ_2 and standard deviation σ_2 (usually unknown).

Notation

- Population 1 has mean μ_1 and standard deviation σ_1 (usually unknown).
- Population 2 has mean μ_2 and standard deviation σ_2 (usually unknown).
- Sample of size n_1 is taken from population 1: the sample mean is $\bar{X}_1$, the sample standard deviation is SD_1, and the standard error is $SE_1 = \frac{SD_1}{\sqrt{n_1}}$

Notation

- Population 1 has mean μ_1 and standard deviation σ_1 (usually unknown).
- Population 2 has mean μ_2 and standard deviation σ_2 (usually unknown).
- Sample of size n_1 is taken from population 1: the sample mean is $\bar{X}_1$, the sample standard deviation is SD_1, and the standard error is $SE_1 = \frac{SD_1}{\sqrt{n_1}}$
- Sample of size n_2 is taken from population 2: the sample mean is $\bar{X}_2$, the sample standard deviation is SD_2, and the standard error is $SE_2 = \frac{SD_2}{\sqrt{n_2}}$

Notation

- Population 1 has mean μ_1 and standard deviation σ_1 (usually unknown).
- Population 2 has mean μ_2 and standard deviation σ_2 (usually unknown).
- Sample of size n_1 is taken from population 1: the sample mean is $\bar{X}_1$, the sample standard deviation is SD_1, and the standard error is $SE_1 = \frac{SD_1}{\sqrt{n_1}}$
- Sample of size n_2 is taken from population 2: the sample mean is $\bar{X}_2$, the sample standard deviation is SD_2, and the standard error is $SE_2 = \frac{SD_2}{\sqrt{n_2}}$
- The two samples are independent.

Estimating the Difference in Means

- of two independent populations: $\mu_1 - \mu_2$

Estimating the Difference in Means

- of two independent populations: $\mu_1 - \mu_2$
- Point estimate $\bar{X}_1 - \bar{X}_2$

Estimating the Difference in Means

- of two independent populations: $\mu_1 - \mu_2$
- Point estimate $\bar{X}_1 - \bar{X}_2$
- Standard Error:

$$SE = \sqrt{(SE_{\bar{x}_1})^2 + (SE_{\bar{x}_2})^2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

Estimating the Difference in Means

- of two independent populations: $\mu_1 - \mu_2$
- Point estimate $\bar{X}_1 - \bar{X}_2$
- Standard Error:

$$SE = \sqrt{(SE_{\bar{x}_1})^2 + (SE_{\bar{x}_2})^2} = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

- 95% CI for $\mu_1 - \mu_2$

$$(\bar{X}_1 - \bar{X}_2) \pm ME$$

where

$$ME = 1.96\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$$

Statistical Significance

- If the 95% confidence interval for $\mu_1 - \mu_2$ **excludes zero (0)**, then we say that the difference is **statistically significant** or that the mean for one group differs significantly than that for the other group.

Statistical Significance

- If the 95% confidence interval for $\mu_1 - \mu_2$ **excludes zero (0)**, then we say that the difference is **statistically significant** or that the mean for one group differs significantly than that for the other group.
- If the 95% confidence interval for $\mu_1 - \mu_2$ **includes zero**, then we say that the difference is **insignificant** or that there is no significant difference between the two population means.

Difference in GPA Averages Example

Let the GPAs today be of group 1 and that of 10 years ago be of group 2.

- point estimate: $\bar{X}_1 - \bar{X}_2 = 2.98 - 2.90 = 0.08$
- standard error:

$$SE = \sqrt{\frac{(.45)^2}{100} + \frac{(.40)^2}{100}} = 0.06$$

- Margin of error:

$$ME = 1.96 \times SE = 1.96 \times 0.06 = 0.118$$

- 95% CI for $\mu_1 - \mu_2$

$$(0.08 - 0.118, 0.08 + 0.118) \Rightarrow (-0.038, 0.198)$$

- The interval **includes zero** and hence, the difference is **insignificant**. Simple chance variability can be a viable explanation for the observed difference.

Diet Comparison Example

- Consider Atkins and Zone diets at 12 months. Denote μ_1 the mean change in BMI for Zone diet group and μ_2 the mean change in BMI for Atkins diet group.

Diet Comparison Example

- Consider Atkins and Zone diets at 12 months. Denote μ_1 the mean change in BMI for Zone diet group and μ_2 the mean change in BMI for Atkins diet group.
- point estimate: $\bar{X}_1 - \bar{X}_2 = (-.53) - (-1.65) = 1.12$

Diet Comparison Example

- Consider Atkins and Zone diets at 12 months. Denote μ_1 the mean change in BMI for Zone diet group and μ_2 the mean change in BMI for Atkins diet group.
- point estimate: $\bar{X}_1 - \bar{X}_2 = (-.53) - (-1.65) = 1.12$
- standard error:

$$SE = \sqrt{\frac{(2.00)^2}{79} + \frac{(2.54)^2}{77}} = 0.367$$

Diet Comparison Example

- Consider Atkins and Zone diets at 12 months. Denote μ_1 the mean change in BMI for Zone diet group and μ_2 the mean change in BMI for Atkins diet group.
- point estimate: $\bar{X}_1 - \bar{X}_2 = (-.53) - (-1.65) = 1.12$
- standard error:

$$SE = \sqrt{\frac{(2.00)^2}{79} + \frac{(2.54)^2}{77}} = 0.367$$

- Margin of error:

$$ME = 1.96 \times SE = 1.96 \times 0.367 = 0.72$$

Diet Comparison Example

- Consider Atkins and Zone diets at 12 months. Denote μ_1 the mean change in BMI for Zone diet group and μ_2 the mean change in BMI for Atkins diet group.
- point estimate: $\bar{X}_1 - \bar{X}_2 = (-.53) - (-1.65) = 1.12$
- standard error:

$$SE = \sqrt{\frac{(2.00)^2}{79} + \frac{(2.54)^2}{77}} = 0.367$$

- Margin of error:

$$ME = 1.96 \times SE = 1.96 \times 0.367 = 0.72$$

- 95% CI for $\mu_1 - \mu_2$

$$(1.12 - 0.72, 1.12 + 0.72) \Rightarrow (0.40, 1.84)$$

Diet Comparison Example

- Consider Atkins and Zone diets at 12 months. Denote μ_1 the mean change in BMI for Zone diet group and μ_2 the mean change in BMI for Atkins diet group.
- point estimate: $\bar{X}_1 - \bar{X}_2 = (-.53) - (-1.65) = 1.12$
- standard error:

$$SE = \sqrt{\frac{(2.00)^2}{79} + \frac{(2.54)^2}{77}} = 0.367$$

- Margin of error:

$$ME = 1.96 \times SE = 1.96 \times 0.367 = 0.72$$

- 95% CI for $\mu_1 - \mu_2$

$$(1.12 - 0.72, 1.12 + 0.72) \Rightarrow (0.40, 1.84)$$

- This CI excludes zero (0) so therefore the difference is statistically significant.

iClicker Question 11.1

A 95% confidence interval for the difference in the means of a numerical measurement on two independent populations was calculated from two independent samples. The result is (1.2, 10.5). Which of the following is true?

1. the confidence interval excludes 0 hence the difference is insignificant
2. the confidence interval includes 0 hence the difference is insignificant
3. the confidence interval includes 0 hence the difference is significant
4. the confidence interval excludes 0 hence the difference is significant

iClicker Question 11.2

A 95% confidence interval for the difference in the means of a numerical measurement on two independent populations was calculated from two independent samples. The result is (1.2, 10.5). Which of the following is a correct interpretation of the confidence interval?

1. we are 95% confident that the difference in sample means is between 1.2 and 10.5
2. we are 95% confident that the difference in the true means is between 1.2 and 10.5
3. there is 95% chance that the difference in the true means is between 1.2 and 10.5
4. there is 95% chance that the difference in sample means is between 1.2 and 10.5
5. none of the above

iClicker Question 11.3

What effect is there on the standard error of the difference in means if the sample sizes are each quadrupled?

1. the standard error is likely to decrease
2. the standard error is likely to increase
3. there is no effect

Computing the P-value for the Difference

in means of two independent populations
For a two-tailed test for the difference in means we can use the z-table to indicate probabilities:

$$\texttt{P-value} = 2 \times \left[1 - P\left[Z \leq \frac{\bar{X}_1 - \bar{X}_2}{SE}\right]\right]$$

Diet Comparison Example, Cont'd

- Is it viable to explain that the difference in changes in BMI of 1.12 is due to chance variability?
- Note that we take the difference in changes in BMI of 1.12 and divide by the SE 0.367 to get a z-value: $1.12/0.367 = 3.05$
- If the true difference were zero (0), the chance to observe a value as far as $\pm$ 3.05 or more SE's away from the mean would be:

Diet Comparison Example, Cont'd

- $$\begin{aligned} P(Z < -3.05 \texttt{ or } Z > 3.05) &= 2P(Z > 3.05) \\ &= 2[1 - P(Z \leq 3.05)] \\ &= 2[1 - 0.9989] \\ &= 0.0022 \end{aligned}$$

- This is a very small chance making it hard to believe that the true difference is zero (0).

- Hence, we conclude that statistically, the two means are different. Or we can say that the means are significantly different.

Statistics and Data Analysis

STAT 1600
Ch 11.2 Comparing Two Means – Paired Data

Outline

Paired Data

- Comparing Means in Paired Data
- (General) Paired Data

Paired Data, Before-and-After Data

Data come in pairs, each subject was measured twice, before and after. Note that it's NOT two samples, it's one sample of subjects, each measured twice. Weight Loss Example:

Table: Weight in pounds before and after 12 months on diet

Subject	Before	After
1	180	155
2	192	187
3	205	194
4	166	176
5	220	205
6	177	172
7	189	173

Analysis of Paired Data

- It's WRONG to analyze the data using two-sample method. The proper procedure is laid out below:
- Calculate the pairwise differences in a consistent manner: $d_i = \texttt{Before} - \texttt{After}, i = 1, \cdots, n$.
- Treat the differences as one sample and
- Mean difference: $\bar{D} = \frac{d_1 + d_2 \cdots d_n}{n}$
- Standard error $SE_{\bar{D}}$:

$$SE = \frac{SD}{n} \;\; \texttt{where} \;\; SD = SD_d = \sqrt{\frac{\sum(d_i - \bar{D})^2}{n-1}}$$

- A 95% CI for μ_d:

$$(\bar{D} \pm ME) \;\; \texttt{where} \;\; ME = 1.96 \times SE$$

Analysis of Paired Data, Cont'd

- If the confidence interval EXCLUDES zero (0), then the difference in means is statistically significant.

Weight Loss Example, revisited

	Before	After	Diff(d)	d - mean
1	180.00	155.00	25.00	15.43
2	192.00	187.00	5.00	-4.57
3	205.00	194.00	11.00	1.43
4	166.00	176.00	-10.00	-19.57
5	220.00	205.00	15.00	5.43
6	177.00	172.00	5.00	-4.57
7	189.00	173.00	16.00	6.43

$$SD = \sqrt{\frac{\sum(d_i - \bar{D})^2}{n-1}} = 11.1$$

Hence,

$$SE = \frac{SD}{\sqrt{n}} = 4.2$$

Weight Loss Example, cont'd

The margin of error is

$$ME = 1.96 \times SE = 1.96 \times 4.2 = 8.2$$

Hence a 95% confidence interval for mean weight loss (i.e., mean difference of Before and After) is

$$((9.6 - 8.2), (9.6 + 8.2)) \Rightarrow (1.4, 17.8)$$

This CI **excludes** zero (0) and hence the mean weight loss is statistically **significant**.

iClicker Question 11.2.1

Measurements of the left-hand and right-hand gripping strengths of 10 left-handed writers are recorded:

	Person									
	1	2	3	4	5	6	7	8	9	10
LH	140	90	125	130	95	121	85	97	131	110
RH	138	87	110	132	96	120	86	90	129	100

Should this data set be treated as a paired data problem?

1. Yes.
2. No.
3. Cannot determine.

iClicker Question 11.2.2

Measurements of the left-hand and right-hand gripping strengths of 10 left-handed writers are recorded:

	Person									
	1	2	3	4	5	6	7	8	9	10
LH	140	90	125	130	95	121	85	97	131	110
RH	138	87	110	132	96	120	86	90	129	100

A 95% confidence interval for $\mu_{left} - \mu_{right}$ is calculated and the result is (0.22,6.98). Which statement below is true about the confidence interval?

1. We are 95% confident that the difference in sample means is between 0.22 and 6.98.
2. There is a 95% probability that the difference in the true mean gripping strengths of the two hands is between 0.22 and 6.98.
3. We are 95% confident that the difference in the true mean gripping strengths is between 0.22 and 6.98.

iClicker Question 11.2.3

Measurements of the left-hand and right-hand gripping strengths of 10 left-handed writers are recorded:

	Person									
	1	2	3	4	5	6	7	8	9	10
LH	140	90	125	130	95	121	85	97	131	110
RH	138	87	110	132	96	120	86	90	129	100

A 95% confidence interval for $\mu_{left} - \mu_{right}$ is calculated and the result is (0.22,6.98). Which statement below is true about the confidence interval?

1. The average gripping strength of the left hand differs significantly than that of the right hand since the confidence interval includes 0.
2. The average gripping strength of the left hand differs significantly than that of the right hand since the confidence interval excludes 0.
3. The difference in average gripping strengths of the two hands is insignificant since the confidence interval includes 0.

(General) Paired Data

- Paired data are data in which natural matchings occur.

(General) Paired Data

- Paired data are data in which natural matchings occur.
- Even when we have two samples, if each observation in one sample is uniquely matched to an observation in the other sample, then we have paired data.

(General) Paired Data

- Paired data are data in which natural matchings occur.
- Even when we have two samples, if each observation in one sample is uniquely matched to an observation in the other sample, then we have paired data.
- The analysis of such data should follow that of before-and-after paired data.

Head Injury Criterion (HIC) Example

	Driver	Passenger
Acura Integra 87	599.00	597.00
Audi 80 89	600.00	515.00
Chev Camaro 91	585.00	583.00
Ford Escort 87	551.00	418.00
Honda Accord LX 91	562.00	539.00
Toyota Corolla Fx 88	593.00	397.00
Volvo 740 LE 88	519.00	445.00

Head Injury Criterion (HIC) eg., Cont'd

- The differences of Head Injury Criterion (HIC) Driver – Passenger: 2, 85, 2, 133, 23, 196, 74

Head Injury Criterion (HIC) eg., Cont'd

- The differences of Head Injury Criterion (HIC) Driver – Passenger: 2, 85, 2, 133, 23, 196, 74
- Mean difference: $\bar{D} = 73.6$

Head Injury Criterion (HIC) eg., Cont'd

- The differences of Head Injury Criterion (HIC) Driver – Passenger: 2, 85, 2, 133, 23, 196, 74
- Mean difference: $\bar{D} = 73.6$
- Variation measures

$$\begin{aligned} SD &= \sqrt{\frac{\sum(d_i - \bar{D})^2}{n-1}} \\ &= 72.4 \end{aligned} \qquad \begin{aligned} SE &= \frac{SD}{\sqrt{n}} \\ &= 27.4 \end{aligned} \qquad \begin{aligned} ME &= 1.96 \times SE \\ &= 53.6 \end{aligned}$$

Head Injury Criterion (HIC) eg., Cont'd

- The differences of Head Injury Criterion (HIC) Driver – Passenger: 2, 85, 2, 133, 23, 196, 74
- Mean difference: $\bar{D} = 73.6$
- Variation measures

$$SD = \sqrt{\frac{\sum(d_i - \bar{D})^2}{n-1}} = 72.4 \qquad SE = \frac{SD}{\sqrt{n}} = 27.4 \qquad ME = 1.96 \times SE = 53.6$$

- A 95% CI for mean difference in HICs (Driver - Passenger):

$$(73.6 - 53.6, 73.6 + 53.6) \Rightarrow (19.9, 127.2)$$

which **excludes** zero (0) and hence the mean difference in HICs is statistically **significant**.

Statistics and Data Analysis

STAT 1600
Ch. 12 Testing Independence of Two Categorical Variables

Outline

Association/Independence of Categorical Variables

- Association/Independence of Categorical Variables
- Testing for Statistical Association

Assoc./Indep. of two Cat. Var.

Association and Independence
Two variables A and B are said to be associated if the distribution of B tends to change with the level of the A variable. Otherwise, they are said to be not associated.
Each of the following are likely to be as stated:

- Gender and height are associated: males tend to be taller than females.

Assoc./Indep. of two Cat. Var.

Association and Independence
Two variables A and B are said to be associated if the distribution of B tends to change with the level of the A variable. Otherwise, they are said to be not associated.
Each of the following are likely to be as stated:

- Gender and height are associated: males tend to be taller than females.
- GPA and height are independent: 'height distribution tends to be the same for 3.0 students as well as 3.5 students.'

Assoc./Indep. of two Cat. Var.

Association and Independence
Two variables A and B are said to be associated if the distribution of B tends to change with the level of the A variable. Otherwise, they are said to be not associated.
Each of the following are likely to be as stated:

- Gender and height are associated: males tend to be taller than females.
- GPA and height are independent: 'height distribution tends to be the same for 3.0 students as well as 3.5 students.'
- Shoe size and height are associated: taller people tend to wear larger-sized shoes.

Example: HS GPA and College Attrition

Is there an association between HS GPA (high school GPA) and college attrition? A sample of $n = 189$ students entering a business school program were followed as part of an attrition (i.e. drop out, transfer) study. The students were cross classified according to Three (3) categories of attrition outcomes and four (4) categories of HS GPA (the Observed frequency table):

	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0
not returned 2nd year	25	3	4	6
not returned 3rd year	14	7	4	6
returned 3rd year	41	7	28	44

GPA and Attrition Example, Cont'd

If grades and attrition were independent, then the (Observed) frequency table should have looked more like the Expected frequency table below:

```
## Warning in chisq.test(MX12A, correct = TRUE):
Chi-squared approximation may be incorrect
```

	2.0-2.5	2.5-3.0	3.0-3.5	3.5-4.0
not returned 2nd year	16.08	3.42	7.24	11.26
not returned 3rd year	13.12	2.79	5.90	9.19
returned 3rd year	50.79	10.79	22.86	35.56

How do we construct such a table?

Constructing Expected Frequency Table

- Compute the row totals of the table.

Constructing Expected Frequency Table

- Compute the row totals of the table.
- Calculate the column totals.

Constructing Expected Frequency Table

- Compute the row totals of the table.
- Calculate the column totals.
- Calculate the grand total which is the sample size n. The total of row totals should equal n and the total of column totals should equal n.

Constructing Expected Frequency Table

- Compute the row totals of the table.
- Calculate the column totals.
- Calculate the grand total which is the sample size n. The total of row totals should equal n and the total of column totals should equal n.
- Calculate the expected frequency counts (these are E's, the expected): For each cell,

$$E_i = \frac{\texttt{row total (i)} \times \texttt{col total (j)}}{\texttt{Grand Total}}$$

iClicker Question 11.2.1

Measurements of the left-hand and right-hand gripping strengths of 10 left-handed writers are recorded:

	Person									
	1	2	3	4	5	6	7	8	9	10
LH	140	90	125	130	95	121	85	97	131	110
RH	138	87	110	132	96	120	86	90	129	100

Should this data set be treated as a paired data problem?

1. Yes.
2. No.
3. Cannot determine.