Inferential Statistics and Probability
a Holistic Approach

Chapter 1
Displaying and Analyzing Data with Graphs

1

1

Introduction – in Class

- Syllabus– Homework 0
- Computer Lab – S44
  - Minitab
- Website
  - http://nebula2.deanza.edu/~mo
  - http://www.professormo.com
- Tutor Lab - S43 (S41 for MPS)
  - Drop in or assigned tutors – get form from lab.
  - Group Tutoring
- Other Questions

2

2

Introduction – Online (Zoom)

- Most material on Canvas
- Install Zoom
- Computer Labs
  - Install Minitab or use Cloud Version
- Website (not Canvas)
  - http://nebula2.deanza.edu/~mo
  - http://www.professormo.com
- Tutor Lab – online
- Drop in or assigned tutors – online.
  - Group Tutoring
- Other Questions

3

3

Descriptive Statistics

- Organizing, summarizing and displaying data
  - Graphs
  - Charts
  - Measure of Center
  - Measures of Spread
  - Measures of Relative Standing

4

4

Problem Solving

- The Role of Probability
- Modeling Random Variables
- Measuring Reliability of an Inference
- Diagnostic Testing
- Simulation of Models
- Verification of Models

5

5

Inferential Statistics

- Population – the set of all measurements of interest to the sample collector
- Sample – a subset of measurements selected from the population
- Inference – A conclusion about the population based on the sample
- Reliability – Measure the strength of the Inference

6

6

Raw Data – Apple
Monthly Adjusted Stock Price: 12/1998 to 4/2020

115.82	102.97	106.17	75.50	69.86	52.70	41.97	27.42	11.11	25.77	11.04	9.35	4.19	1.39	0.93	1.42	0.97
110.52	115.73	114.39	74.83	76.83	49.73	40.49	26.01	12.06	23.71	11.93	8.82	4.36	1.36	1.01	1.39	1.07
112.96	116.40	103.43	69.93	77.80	52.67	39.16	24.53	14.00	24.72	10.55	7.49	3.41	1.49	1.05	1.14	1.27
112.47	107.44	96.49	63.79	87.18	49.62	36.92	24.12	14.79	19.97	10.02	6.98	2.52	1.35	0.94	1.01	1.68
105.56	109.84	98.16	65.19	86.93	50.07	31.63	21.89	22.06	18.02	8.83	6.10	2.24	1.47	0.96	1.21	3.96
103.12	117.62	91.10	60.15	79.47	50.81	33.47	21.26	20.68	17.14	8.84	5.55	2.10	1.37	0.99	1.22	3.31
94.60	121.63	88.56	52.71	75.99	43.68	32.73	18.53	21.79	15.88	7.45	4.79	2.12	1.24	1.15	1.51	3.41
98.81	126.33	86.17	59.78	75.17	45.26	33.43	17.67	24.56	15.77	7.78	5.17	1.83	1.17	1.52	1.30	2.73
92.20	120.85	79.89	58.47	75.99	45.56	33.97	16.37	22.63	12.99	9.16	4.69	1.68	0.93	1.58	1.66	4.04
107.20	120.15	72.66	58.45	78.01	45.35	30.58	13.68	18.67	12.09	8.16	5.42	1.76	0.92	1.54	1.44	4.42
95.10	124.05	71.24	58.28	70.58	45.96	26.63	11.62	16.27	11.01	8.91	5.84	1.56	0.98	1.41	1.19	3.73
95.22	112.69	67.37	59.80	59.40	44.15	24.99	11.73	17.61	11.16	9.83	5.00	1.47	0.93	1.61	1.41	3.38

7

7

Apple – Adjusted Stock Price 20 Years

8

8

Crime Rate

- In the last 18 years, has violent crime:
  - Increased?
  - Stayed about the Same?
  - Decreased?

9

9

Perception – Gallup Poll

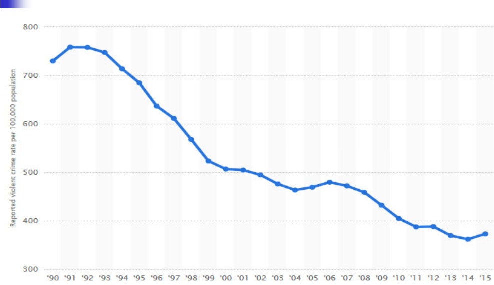
Is there more crime **in the U.S.** than there was a year ago, or less?

	More	Less	Same (vol.)	No opinion
	%	%	%	%
2016 Oct 5-9	70	20	6	4
2015 Oct 7-11	70	18	8	4
2014 Oct 12-15	63	21	9	7
2013 Oct 3-6	64	19	9	7
2011 Oct 6-9	68	17	8	8
2010 Oct 7-10	66	17	8	9
2009 Oct 1-4	74	15	6	5
2008 Oct 3-5	67	15	9	9
2007 Oct 4-7	71	14	8	6
2006 Oct 9-12	68	16	8	8
2005 Oct 13-16	67	21	9	3
2004 Oct 11-14	53	28	14	5
2003 Oct 6-8	60	25	11	4
2002 Oct 14-17	62	21	11	6
2001 Oct 11-14	41	43	10	6
2000 Aug 29-Sep 5	47	41	7	5
1998 Oct 23-25	52	35	8	5
1997 Aug 22-25	64	25	6	5
1996 Jul 25-28	71	15	8	6
1993 Oct 13-18	87	4	5	4
1992 Feb 28-Mar 1	89	3	4	4
1990 Sep 10	84	3	7	6

10

10

Reality - Reported Violent Crime Rate
(Source: US Justice Department)

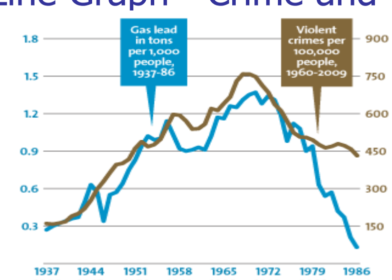


© Statista 2017

11

11

Line Graph - Crime and Lead

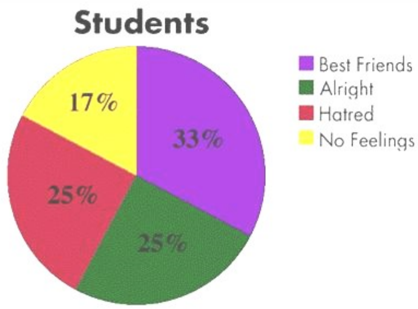


Sources: Rick Nevin, USGS, DOJ

12

12

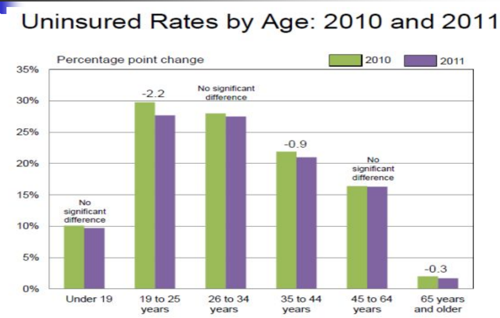
Pie Chart - What do you think of your College roommate?



13

13

Bar Chart - Health Care



14

14

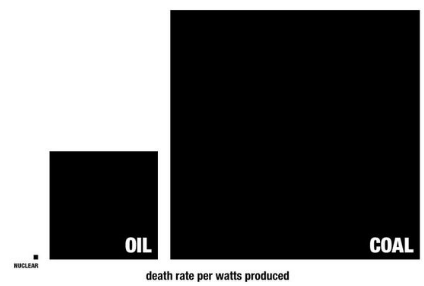
Distorting the truth with Statistics

15

15

Nuclear, Oil and Coal Energy Deaths per terawatt-hour produced

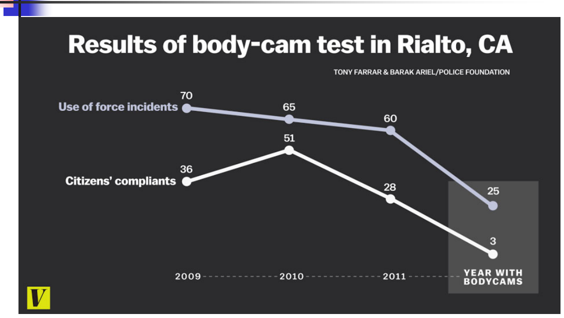
source: thebigfuture.com



16

16

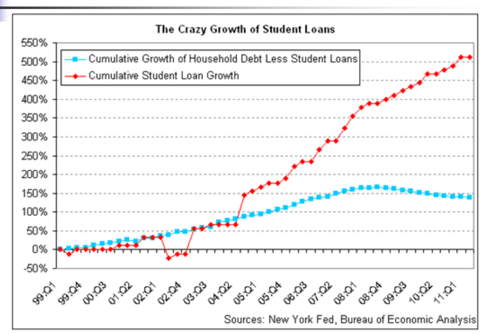
Should Police wear Body Cameras?



17

17

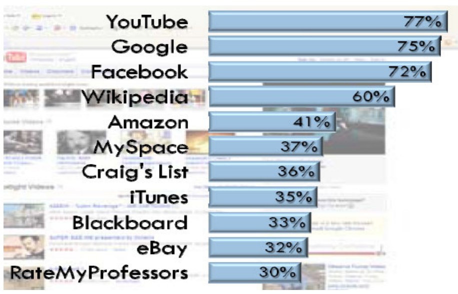
Increase in Debt since 1999



18

18

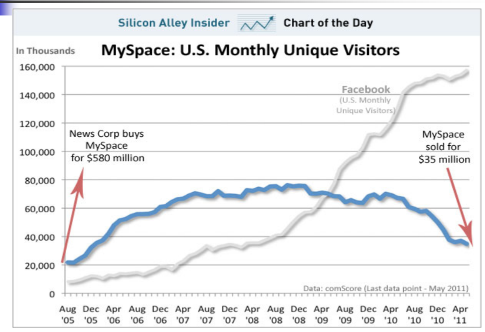
Most Popular Websites for College Students in 2007



19

19

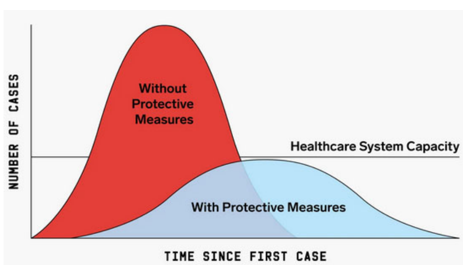
Decline of MySpace



20

20

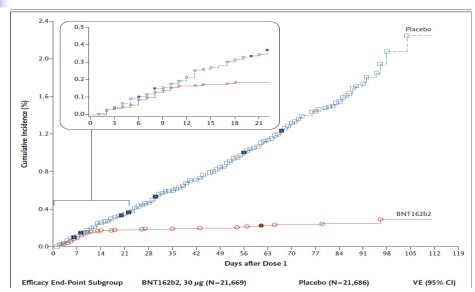
Flattening the Curve



21

21

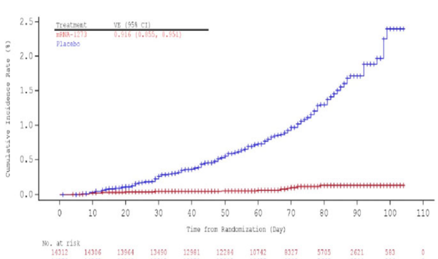
Pfizer Vaccine



22

22

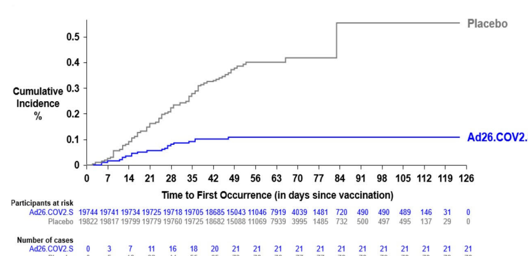
Moderna Vaccine



23

23

Johnson & Johnson Vaccine



24

24

RATE MY PROFESSORS
Over 6,000 Schools, 1 million professors, 6 million opinions

De Anza College

Professor's Name	Department	Total Ratings	Overall Quality	Ease	Hot?
	Mandarin	3	4.3	2.0	
	Mandarin	8	1.6	1.6	
	Marketing	1	5.0	5.0	
	Mathematics	66	4.7	4.0	
	Mathematics	73	1.4	1.7	
	Mathematics	15	2.7	2.6	
	Mathematics	41	1.6	2.1	

25

25

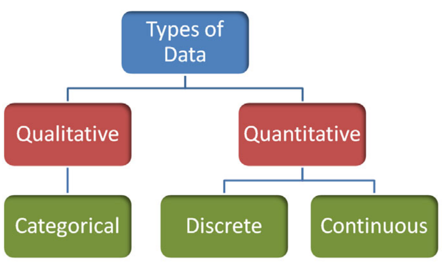
Types of Data

- Qualitative
  - Non-numeric
  - Always categorical
- Quantitative
  - Numeric
  - Categorical numbers are actually qualitative
  - Continuous or discrete

26

26

Types of Data



27

27

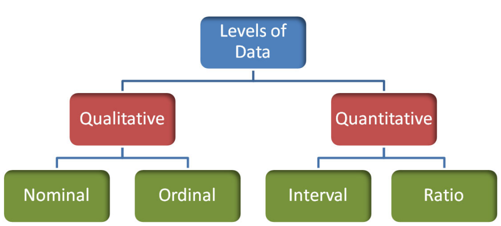
Levels of Data Measurement

- **Nominal:** Names or labels only
  - Example: What city do you live in?
- **Ordinal:** Data can be ranked, but no quantifiable difference.
  - Example: Ratings Excellent, Good, Fair, Poor
- **Interval:** Data can be ranked with quantifiable differences, but no true zero.
  - Example: Temperature
- **Ratio:** Data can be ranked with quantifiable differences and there is a true zero.
  - Example: Age

28

28

Levels of Data Measurement



29

29

Examples of Data

- Distance from De Anza College
- Number of Grandparents still alive
- Eye Color
- Amount you spend on food each week.
- Number of Social Media "Friends"
- Zip Code
- City you live in.
- Year of Birth
- How to prepare Steak? (rare, medium, well-done)
- Do you drive to De Anza?

30

30

Graphical Methods

- Qualitative Data
  - Pie Chart
  - Bar Chart
- Quantitative Data
  - Stem and Leaf Chart
  - Histogram
  - Ogive
  - Dot Plot

31

31

Graphing Categorical Data

A sample of 500 adults (age 18 and over) from Santa Clara County, California were taken from the year 2000 United States Census.

Marital Status	Frequency
Married	270
Widowed	22
Divorced - not remarried	42
Separated	10
Single - never married	156
Total	**500**

32

32

Graphing Categorical Data

- **n = sample size -** The number of observations in your sample size.
- **Frequency -** the number of times a particular value is observed.
- **Relative frequency -** The proportion or percentage of times a particular value is observed.
- **Relative Frequency = Frequency / n**

33

33

Graphing Categorical Data

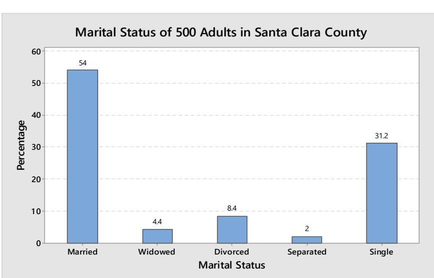
A sample of 500 adults (age 18 and over) from Santa Clara County, California were taken from the year 2000 United States Census.

Marital Status	Frequency	Relative Frequency
Married	270	270/500 = 0.540 or 54.0%
Widowed	22	22/500 = 0.044 or 4.4%
Divorced - not remarried	42	42/500 = 0.084 or 8.4%
Separated	10	10/500 = 0.020 or 2.0%
Single - never married	156	156/500 = 0.312 or 31.2%
Total	**500**	**500/500 = 1.000 or 100.0%**

34

34

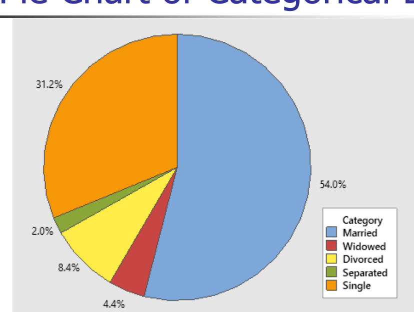
Bar Graph of Categorical Data



35

35

Pie Chart of Categorical Data



36

36

Daily Minutes upload/download on the Internet - 30 students

102	104	85	67	101
71	116	107	99	82
103	97	105	103	95
105	99	86	87	100
109	108	118	87	125
124	112	122	78	92

37

37

Describing Numeric Data

- Center?
  - Where is an "average" value
- Spread?
  - How far are data spread from the center
- Shape?
  - Symmetric or skewed?
- Anything Unusual?
  - Outliers, more than 1 peak?

38

38

Stem and Leaf Graph

```
 6 7
 7 18
 8 25677
 9 25799
10 01233455789
11 268
12 245
```

39

39

Grouping Data

- Choose the number of groups
  - between 5 and 10 is best
- Interval Width = (Range+1)/(Number of Groups)
  - Round **up** to a convenient value
- Start with lowest value and create the groups.
- Example – for 5 categories
  Interval Width = (58+1)/5 = 12 (rounded up)

40

40

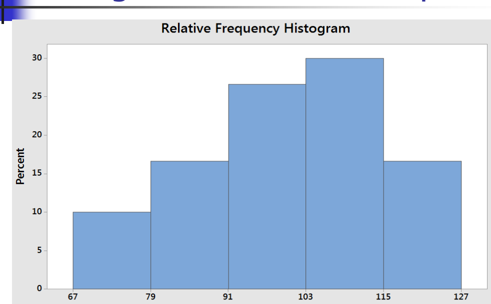
Grouping Data

Class Interval	Frequency	Relative Frequency
67 to 79	3	0.100 or 10.0%
79 to 91	5	0.167 or 16.7%
91 to 103	8	0.266 or 26.6%
103 to 115	9	0.300 or 30.0%
115 to 127	5	0.167 or 16.7%
Total	30	**1.000 or 100%**

41

41

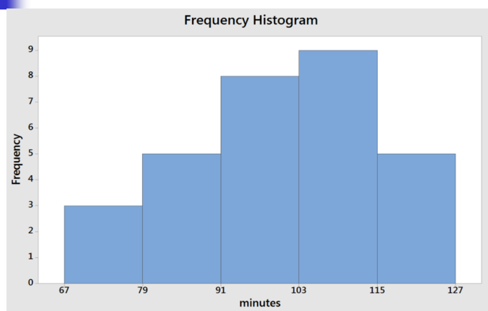
Histogram – Relative Frequency



42

42

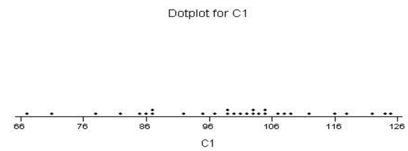
Histogram –Frequency



43

43

Dot Plot – Graph of Frequency



44

44

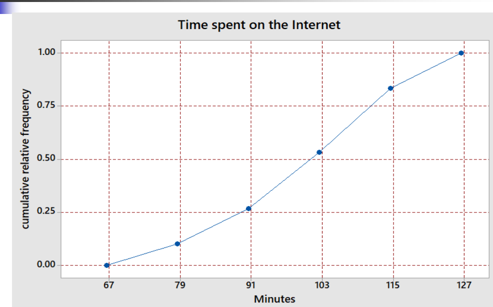
Cumulative Relative Frequency

Class Interval	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
67 to 79	3	0.100 or 10.0%	3	0.100 or 10.0%
79 to 91	5	0.167 or 16.7%	8	0.267 or 26.7%
91 to 103	8	0.266 or 26.6%	16	0.533 or 53.3%
103 to 115	9	0.300 or 30.0%	25	0.833 or 83.3%
115 to 127	5	0.167 or 16.7%	30	1.000 or 100%
Total	30	**1.000 or 100%**		

45

45

Ogive – Graph of Cumulative Relative Frequency



46

46

Inferential Statistics and Probability a Holistic Approach

Chapter 2
Descriptive Statistics

1

1

Example

Anthony's Pizza, a Detroit based company, offers pizza delivery to its customers. A driver for Anthony's Pizza will often make several deliveries on a single delivery run. A sample of 5 delivery runs by a driver showed that the total number of pizzas delivered on each run

2 2 5 9 12

What is the Average?

a) 2
b) 5
c) 6

2

2

Measures of Central Tendency

- Mean
  - Arithmetic Average $\overline{X} = \frac{\sum X_i}{n}$
- Median
  - "Middle" Value after ranking data
  - Not affected by "outliers"
- Mode
  - Most Occurring Value
  - Useful for non-numeric data

3

3

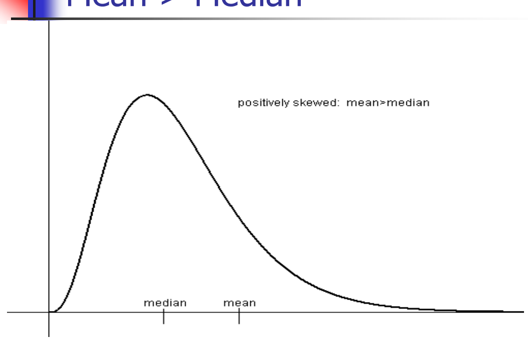
Example – 5 Recent Home Sales

- $500,000
- $600,000
- $600,000
- $700,000
- $2,600,000

4

4

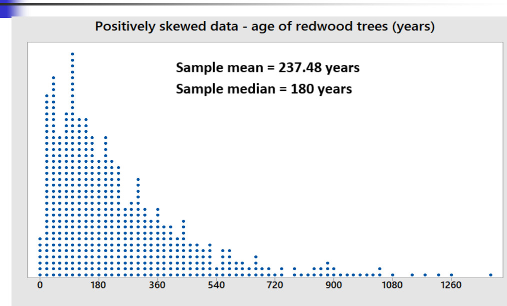
Positively Skewed Data Set Mean > Median



5

5

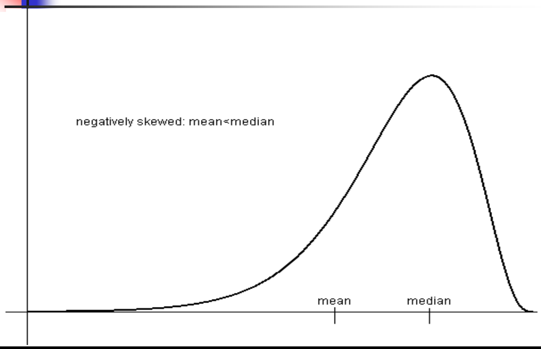
Example – Skewed Positive



6

6

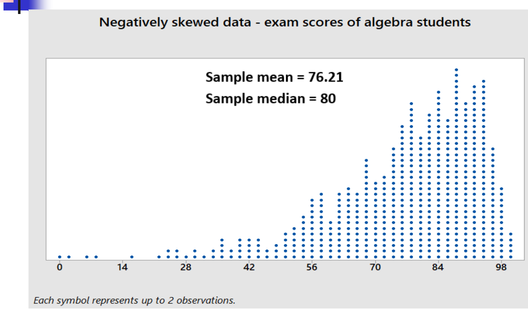
Negatively Skewed Data Set Mean < Median



7

7

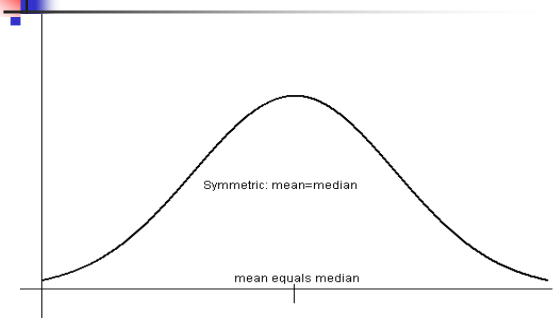
Example – Skewed Negative



8

8

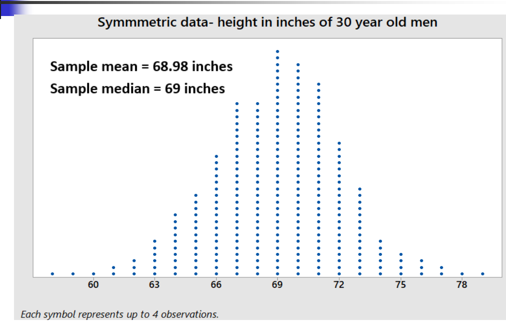
Symmetric Data Set Mean = Median



9

9

Example - Symmetric



10

10

Measures of Variability

- Range
- Variance
- Standard Deviation
- Interquartile Range (percentiles)

11

11

Range

Range = Max(Xi) –Min(Xi) (high – low)

Example – Pizza Delivery
Max = 12 pizzas
Min = 2 pizzas
Range =12 – 2 =10 pizzas

12

12

Sample Variance

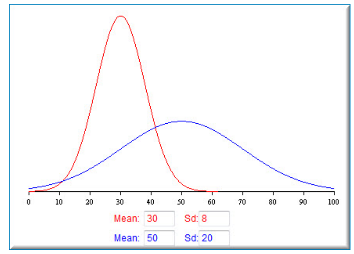
$$s^2 = \frac{Sum\ of\ Squared\ Deviations}{n-1}$$

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

13

13

Sample Standard Deviation



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

14

14

Variance and Standard Deviation

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2		
2		
5		
9		
12		
30		

15

15

Variance and Standard Deviation

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	-4	16
2	-4	16
5	-1	1
9	3	9
12	6	36
30	**0**	**78**

$$s^2 = \frac{78}{4} = 19.5$$

$$s = \sqrt{19.5} \approx 4.42$$

16

16

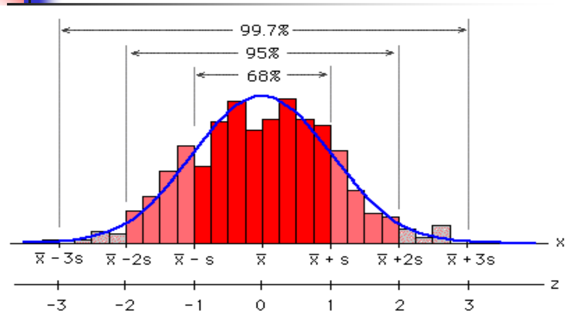
Interpreting the Standard Deviation

- Empirical Rule (68-95-99 rule)
  - For bell shaped data
  - 68% within 1 standard deviation of mean
  - 95% within 2 standard deviations of mean
  - 99.7% within 3 standard deviations of mean

17

17

Empirical Rule

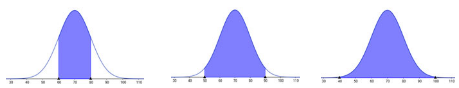


18

18

Example

- An exam has a mean score of 70 and a standard deviation of 10

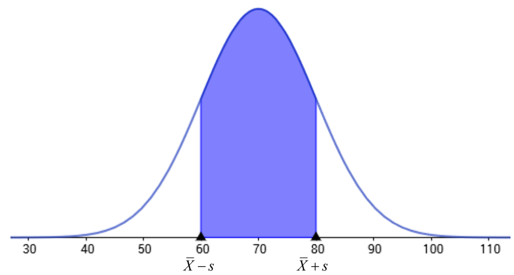


- 68% of scores are between 60 and 80
- 95% of scores are between 50 and 90
- 99.7% of scores are between 40 and 100

19

19

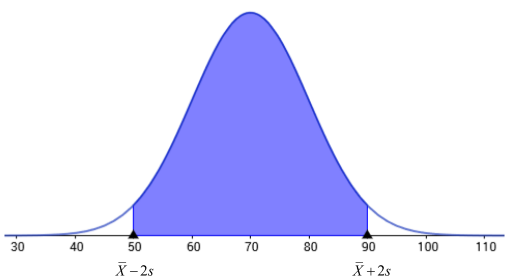
68% of Data



20

20

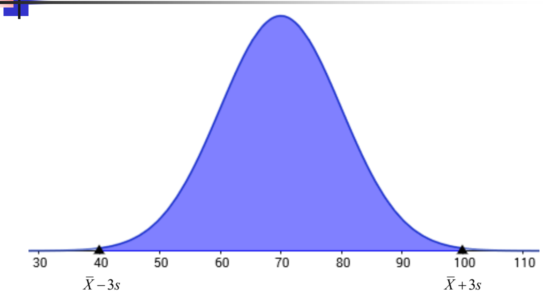
95% of Data



21

21

99.7%% of Data



22

22

Measures of Relative Standing

- Z-score
- Percentile
- Quartiles
- Box Plots

23

23

Z-score

- The number of Standard Deviations from the Mean
- Z>0, X_i is greater than mean
- Z<0, X_i is less than mean

$$Z = \frac{X_i - \bar{X}}{s}$$

24

24

Percentile Rank

Formula for ungrouped data

- The location is (n+1)p (interpolated or rounded)
- n= sample size
- p = percentile

25

25

Quartiles

- 25th percentile is 1st quartile
- 50th percentile is median
- 75th percentile is 3rd quartile
- 75th percentile – 25th percentile is called the Interquartile Range which represents the "middle 50%"

26

26

Alternate method to find Quartiles

- First find median of data. This splits the data into two groups, the lower half and the upper half.
- The median of the lower half of the data is the first quartile.
- The median of the upper half of the data is the third quartile.

27

27

Daily Minutes upload/download on the Internet - 30 students

102	104	85	67	101
71	116	107	99	82
103	97	105	103	95
105	99	86	87	100
109	108	118	87	125
124	112	122	78	92

28

28

Stem and Leaf Graph

```
 6 7
 7 18
 8 25677
 9 25799
10 01233455789
11 268
12 245
```

29

29

IQR Time on Internet data

n+1=31

.25 x 31 = 7.75 location 8 = **87** ← 1st Quartile

.75 x 31 = 23.25 location 23 = **108** ← 3rd Quartile

Interquartile Range (IQR) =108 – 87 = **21**

30

30

Alternate method to find Quartiles

- The median of the data is 101.5
- Q1: The median of the 15 values below 101.5 is 87.
- Q3: The median of the 15 values above 101.5 is 108.
- IQR = 108 - 87 = 21

31

31

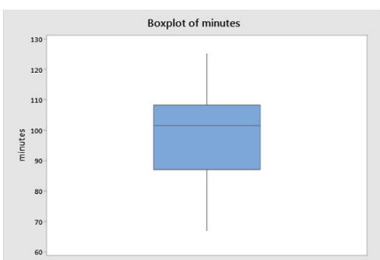
Box Plots

- A box plot is a graphical display, based on quartiles, that helps to picture a set of data.
- Five pieces of data are needed to construct a box plot:
  - Minimum Value
  - First Quartile
  - Median
  - Third Quartile
  - Maximum Value.

32

32

Boxplot



33

33

Outliers

- An outlier is data point that is far removed from the other entries in the data set.
- Outliers could be
  - Mistakes made in recording data
  - Data that don't belong in population
  - True rare events

34

34

Outliers have a dramatic effect on some statistics

- Example quarterly home sales for 10 realtors:

2 2 3 4 5 5 6 6 7 50

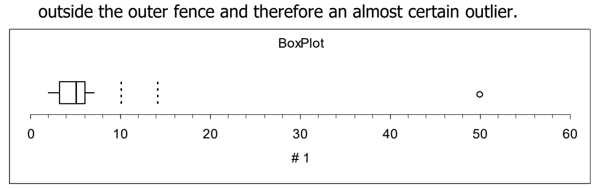
	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

35

35

Using Box Plot to find outliers

- The "box" is the region between the 1st and 3rd quartiles.
- Possible outliers are more than 1.5 IQR's from the box (inner fence)
- Probable outliers are more than 3 IQR's from the box (outer fence)
- In the box plot below, the dotted lines represent the "fences" that are 1.5 and 3 IQR's from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.



36

36

Using Z-score to detect outliers

- Calculate the mean and standard deviation without the suspected outlier.
- Calculate the Z-score of the suspected outlier.
- If the Z-score is more than 3 or less than -3, that data point is a probable outlier.

$$Z = \frac{50 - 4.4}{1.81} = 25.2$$

37

37

Outliers – what to do

- Remove or not remove, there is no clear answer.
- For some populations, outliers don't dramatically change the overall statistical analysis. Example: the tallest person in the world will not dramatically change the mean height of 10000 people.
- However, for some populations, a single outlier will have a dramatic effect on statistical analysis (called "**Black Swan**" by Nicholas Taleb) and inferential statistics may be invalid in analyzing these populations. Example: the richest person in the world will dramatically change the mean wealth of 10000 people.

38

38

Bivariate Data

- Ordered numeric pairs (X,Y)
- Both values are numeric
- Paired by a common characteristic
- Graph as Scatterplot

39

39

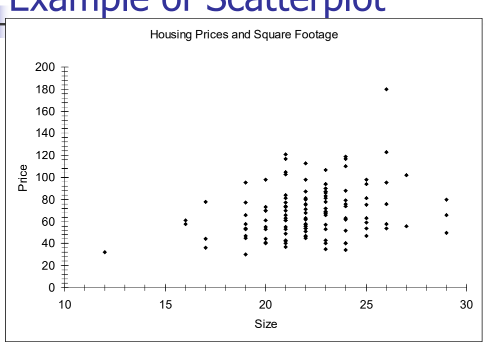
Example of Bivariate Data

- Housing Data
  - X = Square Footage
  - Y = Price

40

40

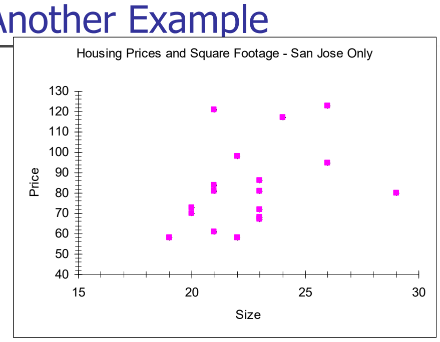
Example of Scatterplot



41

41

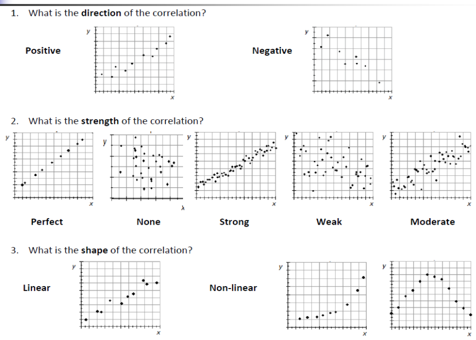
Another Example



42

42

Types of Correlation



43

43

Correlation Analysis

- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated. "Effect"
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable. "Cause?" (Maybe!)

44

44

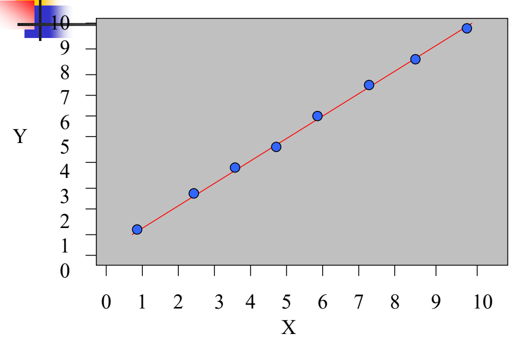
The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the **strength** of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1 to 1.
  - Values of -1 or 1 indicate perfect and strong correlation.
  - Values close to 0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

45

45

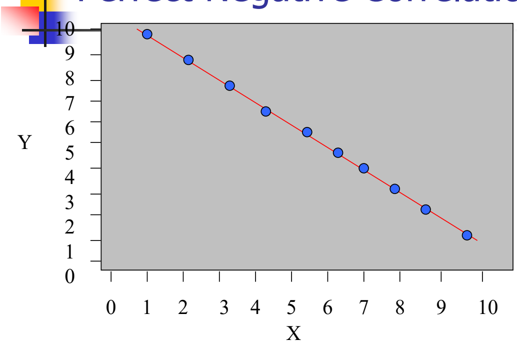
Perfect Positive Correlation



46

46

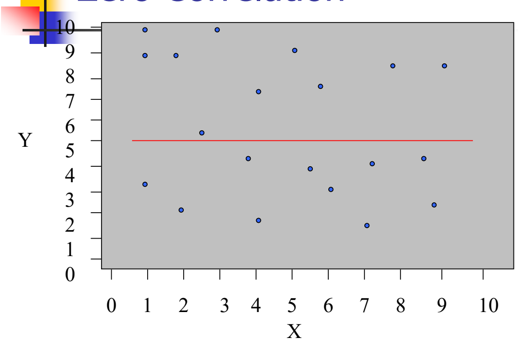
Perfect Negative Correlation



47

47

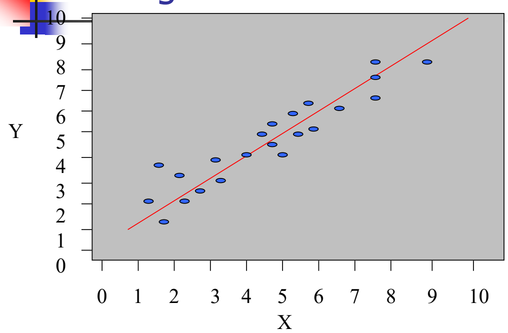
Zero Correlation



48

48

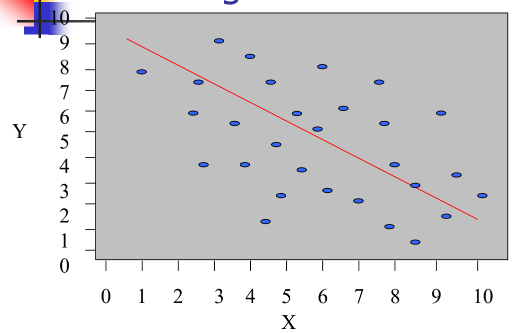
Strong Positive Correlation



49

49

Weak Negative Correlation



50

50

Causation

- Correlation does not necessarily imply causation.
- There are 4 possibilities if X and Y are correlated:
  1. X causes Y
  2. Y causes X
  3. X and Y are caused by something else.
  4. Confounding - The effect of X and Y are hopelessly mixed up with other variables.

51

51

Causation - Examples

- City with more police per capita have more crime per capita.
- As Ice cream sales go up, shark attacks go up.
- People with a cold who take a cough medicine feel better after some rest.

52

52

Formula for correlation coefficient r

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2$$

$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma Y)$$

53

53

Example

- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000
- Make a Scatter Diagram
- Find the correlation coefficient

X	10	15	20	30	40
Y	40	35	25	25	15

54

54

Example *continued*

- Make a Scatter Diagram
- Find the correlation coefficient

55

55

Example *continued*

X	Y	X^2	Y^2	XY
10	40	100	1600	400
15	35	225	1225	525
20	25	400	625	500
30	25	900	625	750
40	15	1600	225	600
115	140	3225	4300	2775

- SSX = $3225 - 115^2/5$ = 580
- SSY = $4300 - 140^2/5$ = 380
- SSXY = 2775 - (115)(140)/5 = -445

56

56

Example *continued*

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}}$$

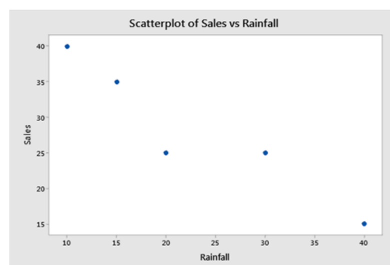
$$r = \frac{-445}{\sqrt{580 \cdot 330}} = -0.9479$$

- Strong negative correlation

57

57

Minitab: Graphs>Scatterplot



58

58

Minitab - Correlation

Stat>
Basic Statistics>
Correlation

Correlations

	Rainfall
Sales	-0.948

59

59

Inferential Statistics and Probability
a Holistic Approach

Chapter 3
Populations and Sampling

1

Population vs. Sample

- A **population** is the entire group of individuals or objects of interest to us.
- A **sample** is a subset of the population that we can study by collecting or gathering data.
- Quantities that describe populations are called **parameters.**
- Quantities that describe samples are called **statistics.**

2

Example

- A large community college has about 25,000 students. In a study of 85 students from college, it was determined that about 60 of the students have moderate or high math anxiety.
- The **population** is **all** the students at this college.
- The **sample** is the 85 students whose math anxiety was measured.

3

Steps of a Statistical Process

- **Step 1 (Problem)**
  Ask a question that can be answered with sample data.
- **Step 2 (Plan)**
  Determine what information is needed.
- **Step 3 (Data)**
  Collect sample data that is representative of the population.
- **Step 4 (Analysis)**
  Summarize, interpret and analyze the sample data.
- **Step 5 (Conclusion)**
  State the results and conclusion of the study.

4

Representative Sample

- A **representative sample** has characteristics, behaviors and attitudes similar to the population from which the sample is selected.
- A sample that is not representative is a **biased sample**.

5

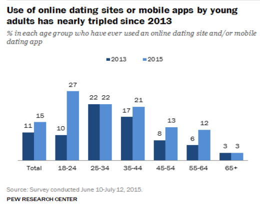
Observational Study

- An **observational study** starts with selecting a representative sample from a population.
- The researcher then takes measurements from the sample, but does not manipulate any of the variables with treatments.
- The goal of an observational study is to interpret and analyze the measured variables, but it is not possible to show a cause and effect relationship.

6

Example of Observational Study

- Pew wanted to investigate a belief that American's use of online dating website and mobile apps had increased from a 2013 study, especially among younger adults.
- A survey was conducted among a national sample of 2,001 adults, 18 years of age or older using random digit dialing
- The survey found that in 2015, 15% of American adults have used online dating sites and mobile apps, compared to 11% in 2013. However, for young adults aged 18-24, the increase was dramatic: from 10% in 2013 to 27% in 2015.



7

7

Example of Observational Study

1: Ask a question that can be answered with sample data.	Has there been an increase in American's use of online dating in the last two years? Are these rates affected by age?
2: Determine what information is needed.	The percentage of adults who are using online dating service. The age of each individual.
3: Collect sample data that is representative of the population.	Since the researchers surveyed both land lines and cell phones using a random dialer, the sample should be representative of the population.
4: Summarize, interpret and analyze the sample data.	15% of American Adults have used online dating sites and mobile apps, compared to 11% in 2013. For young adults aged 18-24, the increase was dramatic: from 10% in 2013 to 27% in 2015. Other age groups are displayed in the graph.
5: State the results and conclusion of the study.	Adults are using online dating sites and mobile dating apps at increasing rates, especially younger adults.

8

8

Experiment

- An **experiment** starts with a representative sample from a population.
- The researcher will them randomly break this sample into groups and then apply treatments in order to manipulate a variable of interest.
- The goal of an experiment is to find a cause and effect relationship between a random variable in the population and the variable manipulated by the researcher.
- If an experiment is conducted properly, the researcher can control for confounding or lurking variables and test for a **placebo effect**.

9

9

Example of Experiment

- Researchers were studying gambling addiction by speed of play using electronic gaming machines.
- 62 participants played a computerized slot machine with either fast, medium, or slow play.
- Gambling speed had no overall effect on either mean bet size, game evaluations or illusion of control, but in the fast machines, at-risk gamblers employed higher bet sizes compared to no-risk gamblers.
- The findings corroborate and elaborate on previous studies and indicate that restrictions on gambling speed may serve as a harm reducing effort for at-risk gamblers.

10

10

Variables in an Experiment

- **Explanatory Variable:** The variable that is controlled or manipulated by the researcher.
- **Response Variable:** The variable which is being measured and is the focus of the study.
- The researcher tries to answer the question: "Does the explanatory variable (cause) affect the response variable (effect)?
- In the prior gambling example, the explanatory variable was the speed of the machine, and the response variable was the bet size.

11

11

Placebos and Blinding

- A **placebo effect** is when participant will respond in a positive way to a treatment with no active ingredients.
- This treatment with no active ingredients is called a **placebo**.
- A **single blind study** is where the participant does not know whether the treatment is real or a placebo.
- A **double blind study** is where neither the administrator of the treatment nor the participant knows whether the treatment is real or a placebo.

12

12

Example

- An researcher for a pharmaceutical company is conducting research on an experimental drug to reduce the pain from migraine headaches.
- Participants with migraine headaches are randomly split into 3 groups. The first group gets the experimental drug (**Treatment Group**). The second group gets a placebo, a fake drug (**Placebo Group**). The third group gets nothing (**Control Group**).
- The researcher found that pain was reduced for both the treatment group and the placebo group, establishing a placebo effect. The researcher must then compare the amount of pain reduction in the treatment group to the placebo group to determine if the treatment was effective.

13

13

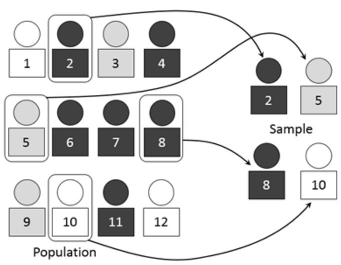
Probability Sampling Methods

- Properly done, probability or scientific sampling will produce a representative sample.
  - Simple Random Sampling
  - Systematic Sampling
  - Stratified Sampling
  - Cluster Sampling

14

14

Simple Random Sampling



15

15

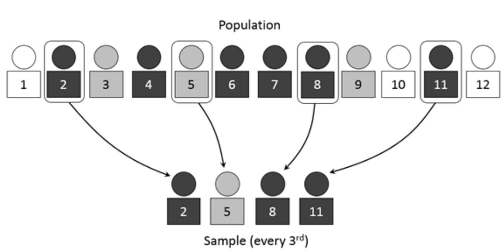
Example - Custom control searching

- Before leaving customs at several international airports, all passengers must push a button.
- If the button is red, you will be required to go through an intensive search.
- If the button is green, you will not be searched
- The button is totally random and has a 20% chance of being red.

16

16

Systematic Sampling



17

17

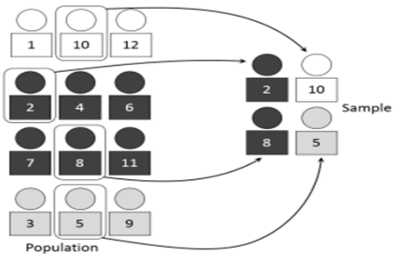
Example – Employee Drug Testing

- A shipping company has approximately 20,000 employees.
- The company decided to administer a random drug test to 5% of the employees, a sample size of 1000.
- The company has a list of all employees sorted by social security number.
- A random number is selected between 1 and 20. Starting with that person, every subsequent 20th person is also sampled.

18

18

Stratified Sampling



19

19

Example – Social Media Black Lives Matter

- Pew Research Center conducted a study to examine how people use social media such as Twitter or Facebook.
- The study focused on the content and hash tags used on people's comments about events involving racially motivated attacks by the police and differences in opinions about groups such as Black Lives Matter.
- Since the study involved people's opinions about race, it was important to use stratified sampling.
- Particular care was taken to make sure that there was appropriate representation in the sample from traditionally under-sampled African American and Latinx groups.

20

20

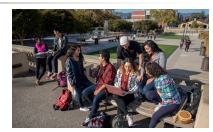
Cluster Sampling

21

21

Example – Sampling Students

- The De Anza College Data Science Club wanted to sample students to complete a survey
- Although the student records are confidential, the classes offered are available to anyone and can be easily sampled
- The club randomly sampled 100 classes offered during the Fall quarter (clusters), and asked the instructors to give the survey link to the students



22

22

Non-Probability Sampling Methods

- Non-probability sampling methods have immeasurable biases and will usually not produce a representative sample.
  - Convenience Sampling
  - Self-selected Sampling

23

23

Example – Convenience Sample

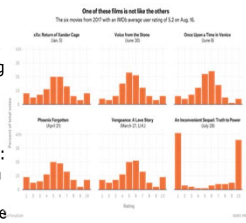
- A 21 year old student wants to conduct a survey on marijuana usage.
- He asks his friends on Facebook to fill out a survey. The results of his survey show that 65% of respondents frequently use marijuana.
- The student's Facebook friends were easy to sample but are not representative of the population.
- For example, if the student frequently uses marijuana, it is more likely that his Facebook friends would also use marijuana

24

24

Example – Self-Selected Sample

- Online rating services, such as Google, Yelp, Rotten Tomatoes, IMDb and Rate My Professors are examples of self-selected sampling
- Users volunteer to write reviews. This can lead to ratings that may be extremely inaccurate.
- Al Gore's "An Inconvenient Sequel: Truth to Power" was released as a follow-up to his original documentary about climate change
- The IMDb rating in this case was not a true movie rating but an attempt to discredit or to support climate change.



25

25

Sources of Bias in Sampling

- **Selection bias** – when the sampling method does not produce a representative sample.
- **Self-selection bias** – when participants who volunteer are not representative of the population.
- **Non-response bias** – when people are intentionally or non-intentionally excluded from participation or choose not to participate in a survey or poll.
- **Response bias** – when the wording of the questions in surveys affect the response.

26

26

Example – Non-response bias

- In 2020 presidential polls understated Trump's Support by about 3.3 percentage points on average, while overstating Biden's support by about 1 point
- This mirrors the polling error in 2016, where Trump's support was also underestimated
- Less educated voters who were a key demographic for Trump on Election Day – are consistently hard for pollsters to reach
- The Rust Belt—which includes Wisconsin, Michigan and Pennsylvania—is notoriously difficult to survey and has many under-sampled Trump supporters

27

27

Example – Response Bias

- Consider these questions:
- "Do you feel that the increasing cost of the high speed rail project is too expensive for California?"
- "Do you feel that high speed rail will be important to the future economy of California?"
- "Do you approve or disapprove of building a high speed rail system in California?"
- The first question encourages people to oppose high speed rail because of the expense.
- The second question encourages people to support high speed rail to support the economy.
- The third question simply asks people's opinion without the leading bias.

28

28

Inferential Statistics and Probability
a Holistic Approach

Chapter 4
Probability

1

1

Probability

- Classical probability
  - Based on mathematical formulas
- Empirical probability
  - Based on the relative frequencies of historical data.
- Subjective probability
  - "one-shot" educated guess.

2

2

Examples of Probability

- What is the probability of rolling a four on a 6-sided die?
- What percentage of De Anza students live in Cupertino?
- What is the chance that your favorite team will win the championship?

3

3

Classical Probability

- Event
  - A result of an experiment, usually expressed as a letter (A, B,…)
- Outcome
  - A result of the experiment that cannot be broken down into smaller events
- Sample Space
  - The set of all possible outcomes
- Probability Event A Occurs - written as P(A)
  - Number of Outcomes in Event A / Number of Outcomes in Sample Space
- Example – flip two coins, find the probability of exactly 1 head.
  - Sample Space = {HH, HT, TH, TT}   A= {HT, TH}
  - P(A) = 2/4 = 0.50

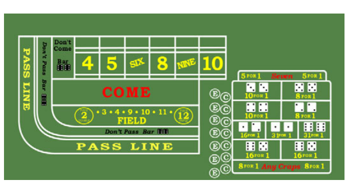
4

4

Example – Field Bet in Craps

Field Bet

- 2 dice are rolled and totaled
- Player wins even money on 3, 4, 9, 10, 11,
- Player wins double on 2
- Player wins triple on 12
- Player loses bet on 5, 6, 7, 8

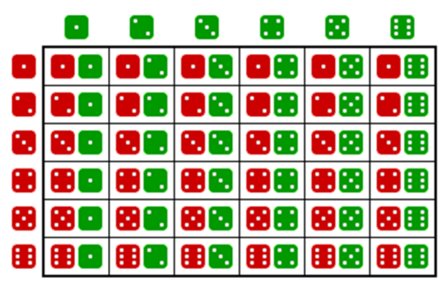


Who has the advantage in this game?

5

5

Sample Space 36 possible pairs of rolls



6

6

More ways to make 5, 6, 7, 8

$$\textit{Sample Space} = \begin{Bmatrix} (1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ (6,1),(6,2),(6,3),(6,4),(6,5),(6,6) \end{Bmatrix}$$

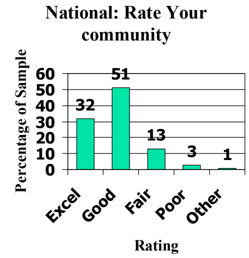
P(Win) = 16/36 = 0.444   P(Lose) = 20/36 = 0.556

7

7

Empirical Probability

- Historical Data
- Relative Frequencies
- Example: What is the chance someone rates their community as good or better?
- **0.51 + 0.32 = 0.83**

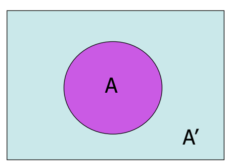


8

8

Rule of Complement

- Complement of an event
- The event does not occur
- A′ is the complement of A
- P(A) + P(A′) =1
- P(A) = 1 – P(A′)

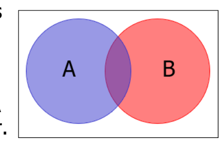


9

9

Joint Probability

- The UNION of two events A and B is that either A or B occur (or both). (All colored parts)
- The INTERSECTION of two events A and B is that both A and B will occur. (Purple Part only)
- Additive Rule:
  P(A or B) = P(A) + P(B) – P(A and B)



10

10

Example

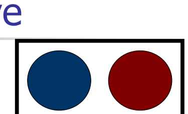
- In a group of students, 40% are taking Math, 20% are taking History.
- 10% of students are taking both Math and History.
- Find the Probability of a Student taking either Math or History or both.
- P(M or H) = 40% + 20% - 10% = 50%

11

11

Mutually Exclusive

- Mutually Exclusive
- Both cannot occur
- If A and B are mutually exclusive, then
  - P(A or B) = P(A) + P(B)
- Example roll a die
  - A: Roll 2 or less   B: Roll 5 or more
  - P(A)=2/6   P(B)=2/6
  - P(A or B) = P(A) + P(B) = 4/6



12

12

Conditional Probability

- The probability of an event occuring GIVEN another event has already occurred.
- P(A|B) = P(A and B) / P(B)
- Example: Of all smart phone users in the US, 21% have an Apple iPhone and AT&T. 35% of all smart phone users have AT&T. Given a selected smart phone user has AT&T, find the probability the user also has an Apple iPhone.
- A=AT&T subscriber B=Apple iPhone
- P(A and B) = 0.21 P(A)=0.35
- P(B|A) = 0.21/0.35 = 0.60

13

13

Marginal, Joint and Conditional Probability

- **Marginal Probability** means the probability of a single event occurring.
- **Joint Probability** means the probability of the union or intersection of multiple events occurring (and/or statements).
- **Conditional Probability** means the probability of an event occurring given that another event has already occurred.

14

14

Contingency Tables

- Two data items can be displayed in a contingency table.
- Example: auto accident during year and DUI of driver.

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000

15

15

Marginal Probability

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000

A = Accident

D = DUI

- Find the Probability a Driver had an Accident
- P(A) = 100/1000 = 0.10
- Find the Probability was not DUI
- P(D') = 1 – 200/1000 = 0.80

16

16

Joint Probability

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000

A = Accident

D = DUI

- Find the Probability a Driver had an Accident **and** was DUI
- P(A and D) = 70/1000 = 0.07
- Find the Probability a Driver had an Accident **or** was DUI
- P(A or D) = P(A) + P(D) – P(A and D) = (100+200-70)/1000 = 0.23

17

17

Conditional Probability

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000

A = Accident

D = DUI

- Find the Probability a DUI Driver had an Accident
- P(A|D) = 70/200 = 0.35
- Find the Probability a Driver who had an Accident is also DUI
- P(D|A) = 70/100 = 0.70

18

18

Creating Contingency Tables

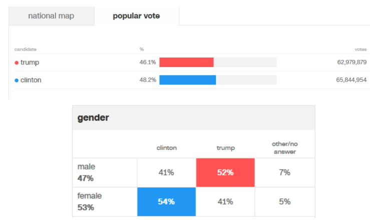
- You can create a hypothetical contingency table from reported cross tabulated data.
- First choose a convenient sample size (called a radix) like 10000.
- Then apply the reported marginal probabilities to the radix of one of the variables.
- Then apply the reported conditional probabilities to the total values of one of the other variable.
- Complete the table with arithmetic.

19

19

Example

Create a two-way table from the cross tabulation of gender from the 2016 election results (from CNN)



20

20

Example

First select a radix (sample size) of 10000

GENDER			
VOTED FOR	Female	Male	Total
Trump			
Clinton			
Other			
Total			10000

21

21

Example

Then apply the marginal probabilities to the radix (53% female, 47% male)

GENDER			
VOTED FOR	Female	Male	Total
Trump			
Clinton			
Other			
Total	5300	4700	10000

22

22

Example

Then apply the cross tabulated percentages for each gender. Make sure the numbers add up.

GENDER			
VOTED FOR	Female	Male	Total
Trump	2173	2444	
Clinton	2862	1927	
Other	265	329	
Total	5300	4700	10000

23

23

Example

Finally, complete the table using arithmetic.

GENDER			
VOTED FOR	Female	Male	Total
Trump	2173	2444	4617
Clinton	2862	1927	4789
Other	265	329	594
Total	5300	4700	10000

24

24

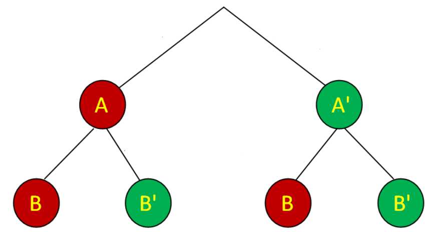
Multiplicative Rule

- P(A and B) = P(A) x P(B|A)
- P(A and B) = P(B) x P(A|B)
- Example: A box contains 4 green balls and 3 red balls. Two balls are drawn. Find the probability of choosing two red balls.
- A=Red Ball on 1st draw   B=Red Ball on 2nd Draw
- P(A)=3/7   P(B|A)=2/6
- P(A and B) = (3/7)(2/6) = 1/7

25

25

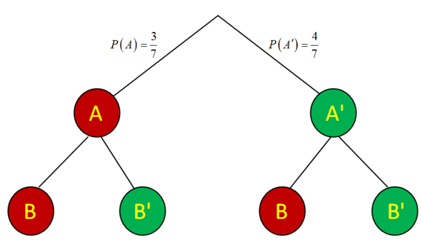
Multiplicative Rule – Tree Diagram



26

26

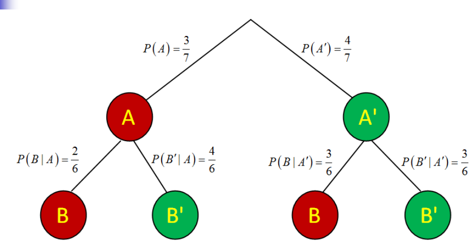
Multiplicative Rule – Tree Diagram



27

27

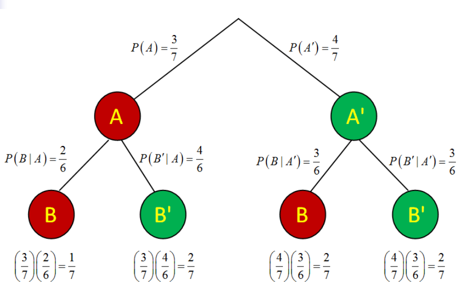
Multiplicative Rule – Tree Diagram



28

28

Multiplicative Rule – Tree Diagram



29

29

Independence

- If A is not dependent on B, then they are INDEPENDENT events, and the following statements are true:
  - P(A|B)=P(A)
  - P(B|A)=P(B)
  - P(A and B) = P(A) x P(B)

30

30

Example

	Accident	No Accident	Total
DUI	70	130	200
Non- DUI	30	770	800
Total	100	900	1000

A: Accident D:DUI Driver

P(A) = .10 P(A|D) = .35 (70/200)

Therefore A and D are **DEPENDENT** events as P(A) < P(A|D)

31

31

Example

	Accident	No Accident	Total
Domestic Car	60	540	600
Import Car	40	360	400
Total	100	900	1000

A: Accident D:Domestic Car

P(A) = .10 P(A|D) = .10 (60/600)

Therefore A and D are **INDEPENDENT** events as P(A) = P(A|D)

Also P(A and D) = P(A)xP(D) = (.1)(.6) = .06

32

32

Random Sample

- A random sample is where each member of the population has an equally likely chance of being chosen, and each member of the sample is INDEPENDENT of all other sampled data.

33

33

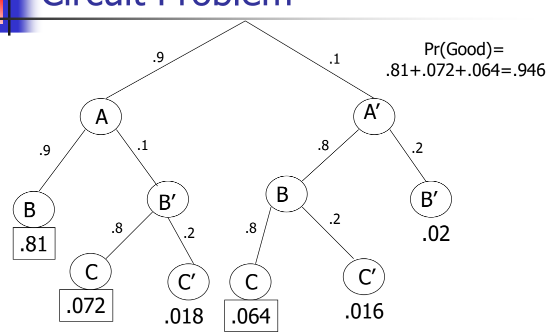
Tree Diagram method

- Alternative Method of showing probability
- Example: Flip Three Coins
- Example: A Circuit has three switches. If at least two of the switches function, the Circuit will succeed. Each switch has a 10% failure rate if all are operating, and a 20% failure rate if one switch has already failed. Find the probability the circuit will succeed.

34

34

Circuit Problem



35

35

Switching the Conditionality

- Often there are questions where you desire to change the conditionality from one variable to the other variable
- First construct a tree diagram.
- Second, create a Contingency Table using a convenient radix (sample size)
- From the Contingency table it is easy to calculate all conditional probabilities.

36

36

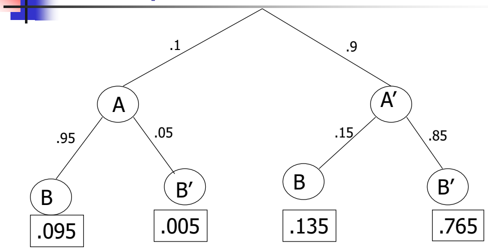
Example

- 10% of prisoners in a Canadian prison are HIV positive.
- A test will correctly detect HIV 95% of the time, but will incorrectly "detect" HIV in non-infected prisoners 15% of the time (false positive).
- If a randomly selected prisoner tests positive, find the probability the prisoner is HIV+

37

37

Example



A=Prisoner is HIV+
B=Test is Positive for HIV

38

38

Example

	HIV+ A	HIV- A′	Total
Test+ B	950	1350	2300
Test- B′	50	7650	7700
Total	1000	9000	10000

$$P(A \mid B) = \frac{950}{2300} \approx .413$$

39

39

Inferential Statistics and Probability
a Holistic Approach

Chapter 5
Discrete Random Variables

1

1

Random Variable

- The value of the variable depends on an experiment, observation or measurement.
- The result is not known in advance.
- For the purposes of this class, the variable will be numeric.

2

2

Random Variables

- Discrete – Data that you Count
  - Defects on an assembly line
  - Reported Sick days
  - RM 7.0 earthquakes on San Andreas Fault
- Continuous – Data that you Measure
  - Temperature
  - Height
  - Time

3

3

Discrete Random Variable

- List Sample Space
- Assign probabilities P(x) to each event x
- Use "relative frequencies"
- Must follow two rules
  - $P(x) \geq 0$
  - $\Sigma P(x) = 1$
- P(x) is called a Probability Distribution Function or pdf for short.

4

4

Probability Distribution Example

- Students are asked 4 questions and the number of correct answers are determined.
- Assign probabilities to each event.

x	P(x)
0	.1
1	.1
2	.2
3	.4
4	

5

5

Probability Distribution Example

- Students are asked 4 questions and the number of correct answers are determined.
- Assign probabilities to each event.

x	P(x)
0	.1
1	.1
2	.2
3	.4
4	.2

6

6

Mean and Variance of Discrete Random Variables

- Population mean μ, is the expected value of x

  $\mu = \Sigma[\ (x)\ P(x)\ ]$

- Population variance σ^2, is the expected value of $(x-\mu)^2$

  $\sigma^2 = \Sigma[\ (x-\mu)^2\ P(x)\ ]$

7

7

Example of Mean and Variance

x	P(x)
0	0.1
1	0.1
2	0.2
3	0.4
4	0.2
Total	**1.0**

8

8

Example of Mean and Variance

x	P(x)	xP(x)
0	0.1	0.0
1	0.1	0.1
2	0.2	0.4
3	0.4	1.2
4	0.2	0.8
Total	**1.0**	**2.5=μ**

9

9

Example of Mean and Variance

x	P(x)	xP(x)	x-μ
0	0.1	0.0	-2.5
1	0.1	0.1	-1.5
2	0.2	0.4	-0.5
3	0.4	1.2	0.5
4	0.2	0.8	1.5
Total	**1.0**	**2.5=μ**	

10

10

Example of Mean and Variance

x	P(x)	xP(x)	x-μ	$(x-\mu)^2$
0	0.1	0.0	-2.5	6.25
1	0.1	0.1	-1.5	2.25
2	0.2	0.4	-0.5	0.25
3	0.4	1.2	0.5	0.25
4	0.2	0.8	1.5	2.25
Total	**1.0**	**2.5=μ**		

11

11

Example of Mean and Variance

x	P(x)	xP(x)	x-μ	$(x-\mu)^2$	$(x-\mu)^2P(x)$
0	0.1	0.0	-2.5	6.25	.625
1	0.1	0.1	-1.5	2.25	.225
2	0.2	0.4	-0.5	0.25	.050
3	0.4	1.2	0.5	0.25	.100
4	0.2	0.8	1.5	2.25	.450
Total	**1.0**	**2.5=μ**			**1.450=σ^2**

12

12

Bernoulli Distribution

- Experiment is one trial
- 2 possible outcomes (Success,Failure)
- p=probability of success
- q=probability of failure
- X=number of successes (1 or 0)
- Also known as Indicator Variable

13

13

Example - Bernoulli

- A basket player makes 70% of free throws. One shot is taken.

- p = P(success) = 0.70
- q = P(failure) = 1-p = 0.30

x	P(x)	xP(x)	(x-μ)	$(x-\mu)^2$	$(x-\mu)^2P(x)$
0	0.30	0	-0.70	0.49	0.147
1	0.70	0.70	0.30	0.09	0.063
Total	**1.0**	**μ=0.70=p**			**σ^2=0.21=pq**

14

14

Mean and Variance of Bernoulli

x	P(x)	xP(x)	$(x-\mu)^2P(x)$
0	(1-p)	0.0	$p^2(1-p)$
1	p	p	$p(1-p)^2$
Total	**1.0**	**p=μ**	**p(1-p)=σ^2**

- $\mu = p$
- $\sigma^2 = p(1-p) = pq$

15

15

Binomial Distribution

- n identical trials
- Two possible outcomes (success/failure)
- Probability of success in a single trial is p
- Trials are mutually independent
- X is the number of successes

- Note: X is a sum of n independent identically distributed Bernoulli distributions

16

16

Example - Binomial

- A basket player makes 70% of free throws. three shots are taken. Find the probability of making exactly 2 Shots

- p = P(Success) = 0.70
- q = P(Failure) = 1-p = 0.30
- n = number of independent trials =3

- P(SSF) = P(S)P(S)P(F) = (0.70)(0.70)(0.30) = 0.147
- P(SFS) = P(S)P(F)P(S) = (0.70)(0.30)(0.70) = 0.147
- P(FSS) = P(F)P(S)P(S) = (0.30)(0.70)(0.70) = 0.147

- P(X=2) = 3(0.147)= 0.441

17

17

Binomial Distribution

- n independent Bernoulli trials
- Mean and Variance of Binomial Distribution is just sample size times mean and variance of Bernoulli Distribution

$$p(x) = {}_nC_x p^x (1-p)^{n-x}$$

$$\mu = E(X) = np$$

$$\sigma^2 = Var(X) = np(1-p)$$

18

18

Binomial Examples

- The number of defective parts in a fixed sample.
- The number of adults in a sample who support the war in Iraq.
- The number of correct answers if you guess on a multiple choice test.

19

19

Binomial Example

- 90% of super duplex globe valves manufactured are good (not defective). A sample of 10 is selected.
- Find the probability of exactly 8 good valves being chosen.
- Find the probability of 9 or more good valves being chosen.
- Find the probability of 8 or less good valves being chosen.

20

20

Using Technology

X	p(X)	cumulative probability
0	0.00000	0.00000
1	0.00000	0.00000
2	0.00000	0.00000
3	0.00001	0.00001
4	0.00014	0.00015
5	0.00149	0.00163
6	0.01116	0.01280
7	0.05740	0.07019
8	0.19371	0.26390
9	0.38742	0.65132
10	0.34868	1.00000
	1.00000	

9.000 expected value
0.900 variance
0.949 standard deviation

Use Minitab or Excel to make a table of Binomial Probabilities.

P(X=8) = .19371

P(X≤8) = .26390

P(X≥9) = 1 - P(X≤8) = .73610

21

21

Poisson Distribution

- Occurrences per time period (rate)
- Rate (μ) is constant
- No limit on occurrences over time period

$$P(x) = \frac{e^{-\mu}\mu^{x}}{x!}$$

$$\mu = \mu$$

$$\sigma = \sqrt{\mu}$$

22

22

Examples of Poisson

- Text messages in the next hour
- Earthquakes on a fault
- Customers at a restaurant
- Flaws in sheet metal produced
- Lotto winners

Note: A binomial distribution with a large n and small p is approximately Poisson with $\mu \approx np$.

23

23

Poisson Example

- Earthquakes of Richter magnitude 3 or greater occur on a certain fault at a rate of twice every year.
- Find the probability of at least one earthquake of RM 3 or greater in the next year.

$$P(X > 0) = 1 - P(0)$$

$$= 1 - \frac{e^{-2}2^{0}}{0!}$$

$$= 1 - e^{-2} \approx .8647$$

24

24

Poisson Example (cont)

- Earthquakes of Richter magnitude 3 or greater occur on a certain fault at a rate of twice every year.
- Find the probability of exactly 6 earthquakes of RM 3 or greater in the next **2** years.

$$\mu = 2(2) = 4$$

$$P(X = 6) = \frac{e^{-4} 4^{6}}{6!} \approx .1042$$

25

25

Inferential Statistics and Probability
a Holistic Approach

Chapter 6
Continuous Random Variables

1

1

Continuous Distributions

- "Uncountable" Number of possibilities
- Probability of a point makes no sense
- Probability is measured over intervals
- Comparable to Relative Frequency Histogram – Find Area under curve.

2

2

Discrete vs Continuous

- Countable
- Discrete Points
- p(x) is probability distribution function
- $p(x) \geq 0$
- $\Sigma p(x) = 1$

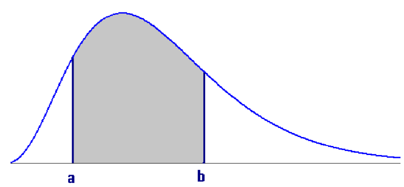
- Uncountable
- Continuous Intervals
- f(x) is probability density function
- $f(x) \geq 0$
- Total Area under curve =1

3

3

Continuous Random Variable

- f(x) is a density function
- P(X<x) is a distribution function.
- P(a<X<b) = area under function between a and b

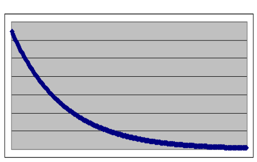


4

4

Exponential distribution

- Waiting time
- "Memoryless"
- $f(x) = (1/\mu)e^{-(1/\mu)x}$
- $P(x>a) = e^{-(a/\mu)}$
- $\mu=\mu \quad \sigma^2=\mu^2$
- $P(x>a+b|x>b) = e^{-(a/\mu)}$



5

5

Examples of Exponential Distributiuon

- Time until…
- a circuit will fail
- the next RM 7 Earthquake
- the next customer calls
- An oil refinery accident
- you buy a winning lotto ticket

6

6

Relationship between Poisson and Exponential Distributions

- If occurrences follow a Poisson Process with mean = μ, then the waiting time for the next occurrence has Exponential distribution with mean = $1/\mu$.
- Example: If accidents occur at a plant at a constant rate of 3 per month, then the expected waiting time for the next accident is 1/3 month.

7

7

Exponential Example

The time until a screen is cracked on a smart phone has exponential distribution with μ=500 hours of use.

(a) Find the probability screen will not crack for at least 600 hours.

$P(x>600) = e^{-600/500} = e^{-1.2} = .3012$

(b) Assuming that screen has already lasted 500 hours without cracking, find the chance the display will last an additional 600 hours.

$P(x>1100|x>500) = P(x>600) = .3012$

8

8

Exponential Example

The time until a screen is cracked on a smart phone has exponential distribution with μ=500 hours of use.

(a) Find the median of the distribution

$P(x>med) = e^{-(med)/500} = 0.5$

med = -500ln(.5) =347

p^{th} Percentile = $-\mu \ln(1-p)$

9

9

Uniform Distribution

- Rectangular distribution
- Example: Random number generator

$$f(x) = \frac{1}{b-a} \quad a \le x \le b$$

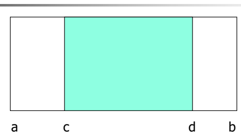
$$\mu = E(X) = \frac{b+a}{2}$$

$$\sigma^2 = Var(X) = \frac{(b-a)^2}{12}$$

10

10

Uniform Distribution - Probability

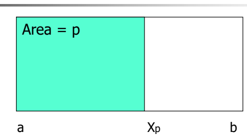


$$P(c < X < d) = \frac{d-c}{b-a}$$

11

11

Uniform Distribution - Percentile



Formula to find the pth percentile X_p:

$$X_p = a + p(b-a)$$

12

12

Uniform Example 1

- Find mean, variance, P(X<3) and 70th percentile for a uniform distribution from 1 to 11.

$$\mu = \frac{1+11}{2} = 6 \quad \sigma^2 = \frac{(11-1)^2}{12} = 8.33$$

$$P(X<3) = \frac{3-1}{11-1} = 0.3$$

$$X_{70} = 1 + 0.7(11-1) = 8$$

13

13

Uniform Example 2

- A tea lover orders 1000 grams of Tie Guan Yin loose leaf when his supply gets to 50 grams.
- The amount of tea currently in stock follows a uniform random variable.
- Determine this model
- Find the mean and variance
- Find the probability of at least 700 grams in stock.
- Find the 80th percentile

14

14

Uniform Example 3

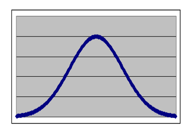
- A bus arrives at a stop every 20 minutes.
  - Find the probability of waiting more than 15 minutes for the bus after arriving randomly at the bus stop.
  - If you have already waited 5 minutes, find the probability of waiting an additional 10 minutes or more. (Hint: recalculate parameters a and b)

15

15

Normal Distribution

- The normal curve is *bell-shaped*
- The mean, median, and mode of the distribution are equal and located at the peak.
- The normal distribution is symmetrical about its mean. Half the area under the curve is above the peak, and the other half is below it.
- The normal probability distribution is asymptotic - the curve gets closer and closer to the x-axis but never actually touches it.

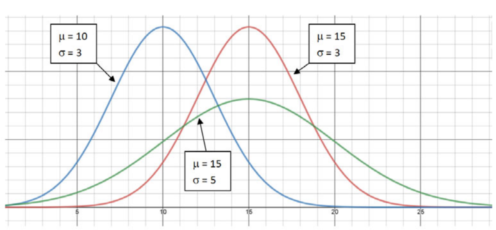


$$f(x) = \frac{e^{-\frac{1}{2\sigma}(x-\mu)^2}}{\sigma\sqrt{2\pi}}$$

16

16

Examples of Normal Random Variables



17

17

The Standard Normal Probability Distribution

- A normal distribution with a mean of 0 and a standard deviation of 1 is called the standard normal distribution.
- Z value: The distance between a selected value, designated x, and the population mean μ, divided by the population standard deviation, σ

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

18

18

Areas Under the Normal Curve – Empirical Rule

- About 68 percent of the area under the normal curve is within one standard deviation of the mean. $\mu \pm 1\sigma$
- About 95 percent is within two standard deviations of the mean $\mu \pm 2\sigma$
- 99.7 percent is within three standard deviations of the mean. $\mu \pm 3\sigma$

19

19

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- About 68% of the daily water usage per person in New Providence lies between what two values?
- $\mu \pm 1\sigma = 20 \pm 1(5)$. That is, about 68% of the daily water usage will lie between 15 and 25 gallons.

20

20

Normal Distribution – probability problem procedure

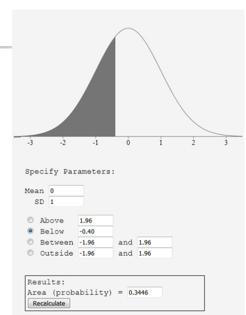
- Given: Interval in terms of X
- Convert to Z by $Z = \dfrac{X - \mu}{\sigma}$
- Look up probability in table or technology

21

21

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- What is the probability that a person from the town selected at random will use less than 18 gallons per day?
- The associated Z value is Z=(18-20)/5=0.
- P(X<18)=P(Z<-0.40)=**0.3446**

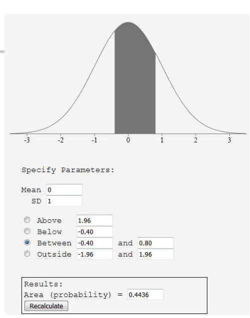


22

22

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- What proportion of the people uses between 18 and 24 gallons?
- The Z value associated with x=18, Z = (18-20)/5 = -0.40 x=24, Z = (24-20)/5 = 0.80.
- P(18<X<24)= P(-0.40<Z<0.80) = .7881. – 3446 = **0.4435**

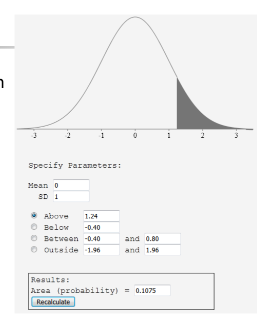


23

23

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons.
- What percentage of the population uses more than 26.2 gallons?
- The Z value associated with X=26.2, Z=(26.2-20)/5=1.24.
- P(X>26.2)=P(Z>1.24) =1-.8925= **0.1075**



24

24

Normal Distribution – percentile problem procedure

- Given: probability or percentile desired.
- Use table or technology that corresponds to probability to get Z
- Convert to X by the formula:

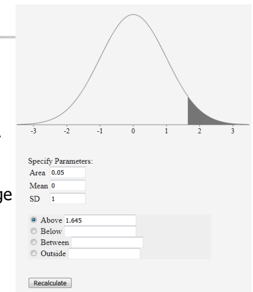
$$X = \mu + Z\sigma$$

25

25

EXAMPLE

- The daily water usage per person in a town is normally distributed with a mean of 20 gallons and a standard deviation of 5 gallons. A special tax is going to be charged on the top 5% of water users.
- Find the value of daily water usage that generates the special tax
- The Z value associated with 95th percentile =1.645
- X=20 + 5(1.645)
  = **28.2 gallons per day**

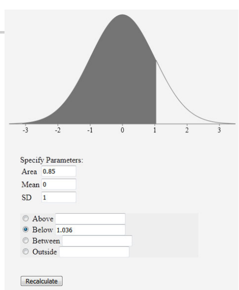


26

26

EXAMPLE

- Professor Kurv has determined that the final averages in his statistics course is normally distributed with a mean of 77.1 and a standard deviation of 11.2.
- He decides to assign his grades for his current course such that the top 15% of the students receive an A.
- What is the lowest average a student can receive to earn an A?
- The top 15% would be the finding the 85th percentile.
- The corresponding Z value is 1.036
- Thus we have X=77.1+(1.036)(11.2), or **X=88.7**

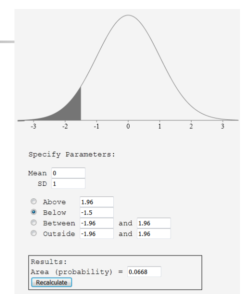


27

27

EXAMPLE

- The amount of tip the servers in an exclusive restaurant receive per shift is normally distributed with a mean of $80 and a standard deviation of $10.
- Shelli feels she has had a bad shift if her total tip for the shift is less than $65.
- What percentage of the time will she feel like she provided poor service?
- Let y be the amount of tip. X=65, Z= (65-80)/10= -1.5.
- P(X<65)=P(Z<-1.5)=**0.0668**



28

28

Inferential Statistics and Probability
a Holistic Approach

Chapter 7
Central Limit Theorem

1

1

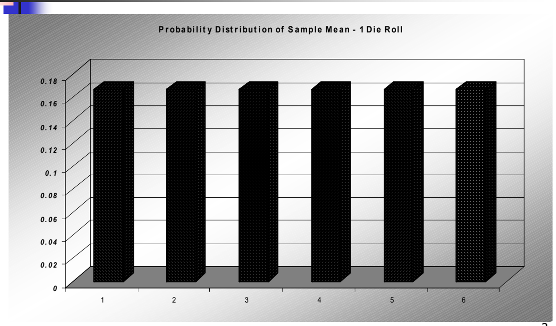
Distribution of Sample Mean

- Random Sample: $X_1, X_2, X_3, \ldots, X_n$
  - Each X_i is a Random Variable from the same population
  - All X_i's are Mutually Independent
- $\overline{X}$ is a function of Random Variables, so $\overline{X}$ is itself Random Variable.
- In other words, the Sample Mean can change if the values of the Random Sample change.
- What is the Probability Distribution of $\overline{X}$?

2

2

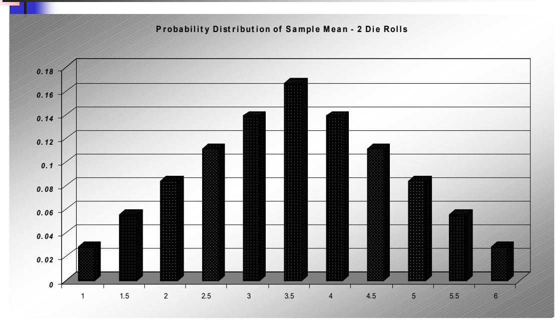
Example – Roll 1 Die



3

3

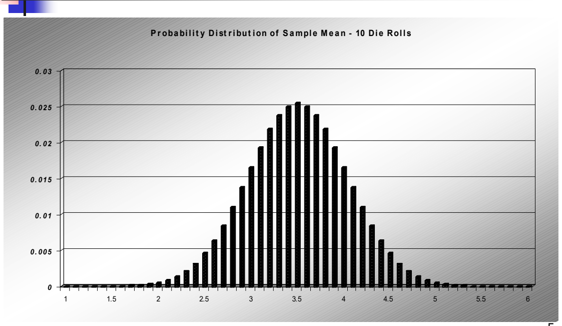
Example – Roll 2 Dice



4

4

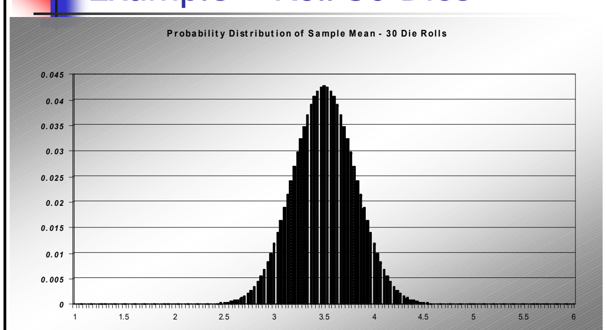
Example – Roll 10 Dice



5

5

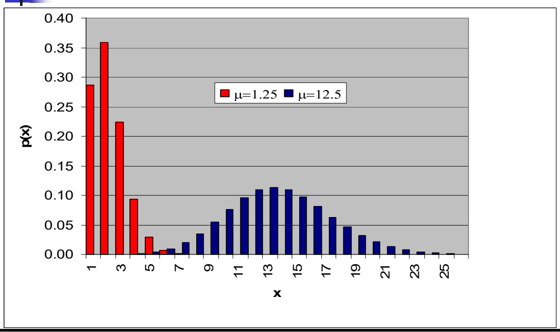
Example – Roll 30 Dice



6

6

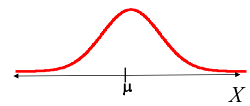
Example - Poisson



7

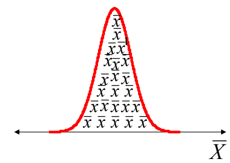
Central Limit Theorem – Part 1

- IF a Random Sample of **any size** is taken from a population with a **Normal Distribution** with mean= μ and standard deviation = σ



- THEN the distribution of the sample mean has a Normal Distribution with:

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

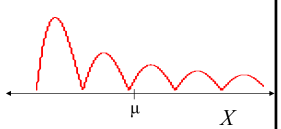


8

8

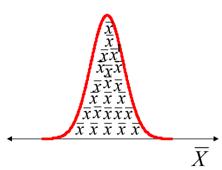
Central Limit Theorem – Part 2

- IF a random sample of **sufficiently large size** is taken from a population with **any Distribution** with mean= μ and standard deviation = σ



- THEN the distribution of the sample mean has approximately a Normal Distribution with:

$$\mu_{\overline{X}} = \mu \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$



9

9

Central Limit Theorem

3 important results for the distribution of $\overline{X}$

- Mean Stays the same

$$\mu_{\overline{X}} = \mu$$

- Standard Deviation Gets Smaller

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

- If n is sufficiently large, $\overline{X}$ has a Normal Distribution

10

10

Example

The mean height of American men (ages 20-29) is μ = 69.2 inches. If a random sample of 60 men in this age group is selected, what is the probability the mean height for the sample is greater than 70 inches? Assume σ = 2.9".

$$P(\overline{X} > 70) = P\left(Z > \frac{(70 - 69.2)}{2.9/\sqrt{60}} \right)$$

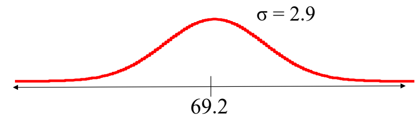
$$= P(Z > 2.14) = 0.0162$$

11

11

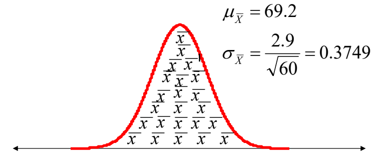
Example (cont)

μ = 69.2
σ = 2.9



$\mu_{\overline{X}} = 69.2$

$\sigma_{\overline{X}} = \frac{2.9}{\sqrt{60}} = 0.3749$



12

12

Example – Central Limit Theorem

The waiting time until receiving a text message follows an exponential (skewed) distribution with an expected waiting time of 1.5 minutes. Find the probability that the mean waiting time for the 50 text messages exceeds 1.6 minutes.

$\mu = 1.5 \qquad \sigma = 1.5 \qquad n = 50$

Use Normal Distribution (n>30)

$$P(\overline{X} > 1.6) = P\left(Z > \frac{(1.6-1.5)}{1.5/\sqrt{50}}\right) = P(Z > 0.47) = 0.3192$$

13

13

Binomial Distribution and Sample Proportion

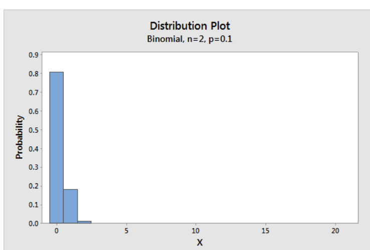
- Let X have a binomial distribution
  - n independent trials
  - p is the probability of success on a single trial
  - X is the number of successes in sample
- Sample proportion
  - $\hat{p}$ is the proportion of successes in sample

$$\hat{p} = \frac{X}{n}$$

14

14

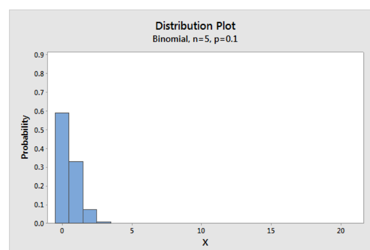
Binomial np=0.2



15

15

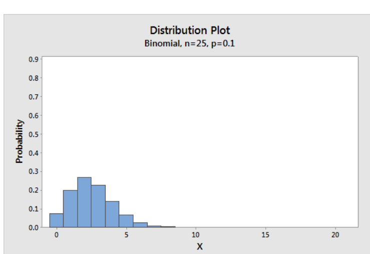
Binomial np=0.5



16

16

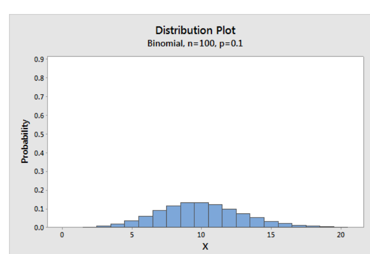
Binomial np=2.5



17

17

Binomial np=10



18

18

Central Limit Theorem
Sample Proportion

- The sample proportion of successes from a sample from a Binomial distribution is a random variable.
- If X is a random variable from a Binomial distribution with parameters n and p,
  and np ≥ 10 and n(1-p) ≥ 10,
  then the following is true for the Sample Proportion, $\hat{P}$:
- $$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$
- The Distribution of $\hat{P}$ is approximately Normal.

19

19

Example

- 45% of all community college students in California receive fee waivers.
- Suppose you randomly sample 1000 community college students to determine the proportion of students with fee waivers in the sample.
- 483 of the sampled students are receiving fee waivers.
- Determine $\hat{P}$. Is the result unusual?

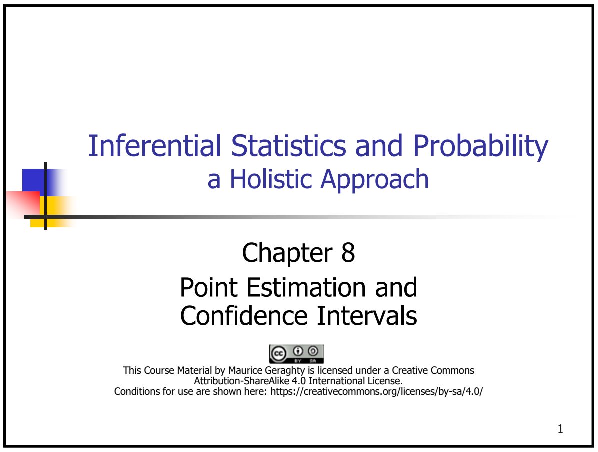
$$\hat{P} = \frac{483}{1000} = 0.483 \qquad \sigma_{\hat{p}} = \sqrt{\frac{0.45(1-0.45)}{1000}} = 0.0157 \qquad Z = \frac{\hat{p}-p}{\sigma_{\hat{p}}} = \frac{0.483-0.45}{0.0157}$$

$$Z = 2.10$$

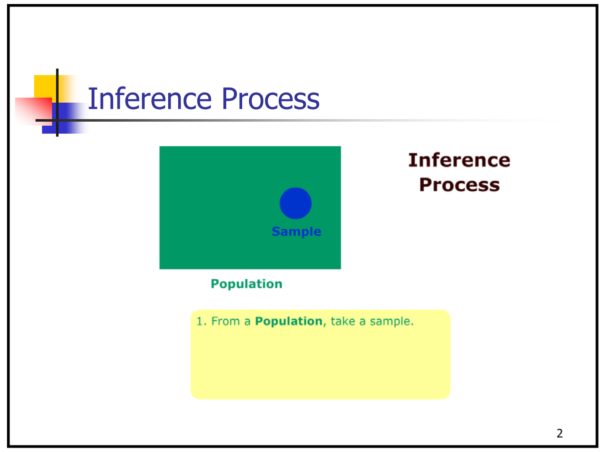
- Result is unusual (more than 2 standard deviations from the expected value of the sample proportion).

20

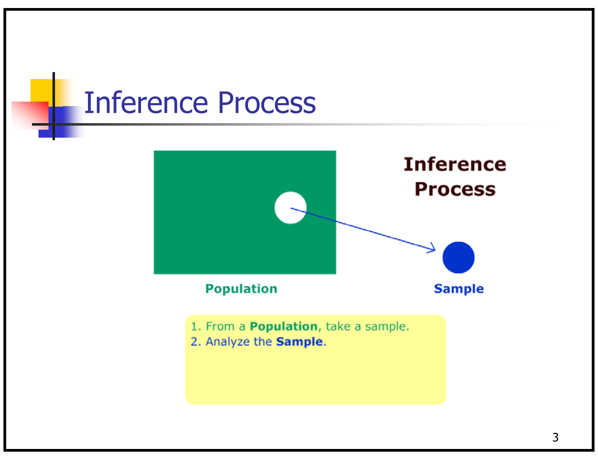
20


Inferential Statistics and Probability
a Holistic Approach
Chapter 8
Point Estimation and
Confidence Intervals
This Course Material by Maurice Geraghty is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Conditions for use are shown here: https://creativecommons.org/licenses/by-sa/4.0/
1

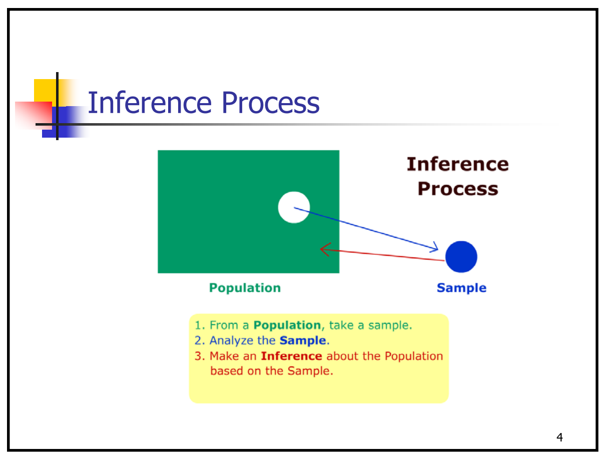
1


Inference Process
Inference Process
Sample
Population
1. From a Population, take a sample.
2

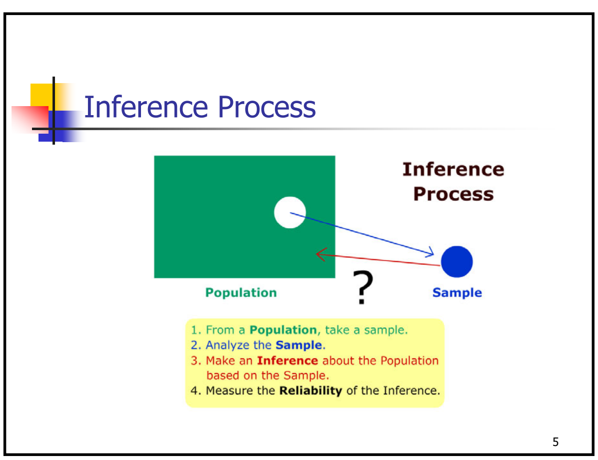
2


Inference Process
Inference Process
Population
Sample
1. From a Population, take a sample.
2. Analyze the Sample.
3

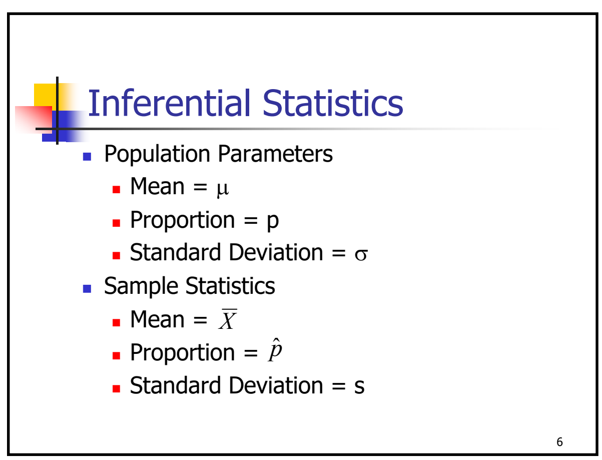
3


Inference Process
Inference Process
Population
Sample
1. From a Population, take a sample.
2. Analyze the Sample.
3. Make an Inference about the Population based on the Sample.
4

4


Inference Process
Inference Process
Population
?
Sample
1. From a Population, take a sample.
2. Analyze the Sample.
3. Make an Inference about the Population based on the Sample.
4. Measure the Reliability of the Inference.
5

5


Inferential Statistics
Population Parameters
Mean = μ
Proportion = p
Standard Deviation = σ
Sample Statistics
Mean = $\bar{X}$
Proportion = $\hat{p}$
Standard Deviation = s
6

6

Inferential Statistics

- Estimation
  - Using sample data to estimate population parameters.
  - Example: Public opinion polls
- Hypothesis Testing
  - Using sample data to make decisions or claims about population
  - Example: A drug effectively treats a disease

7

7

Estimation of μ

$\overline{X}$ is an unbiased point estimator of μ

Example: The number of defective items produced by a machine was recorded for five randomly selected hours during a 40-hour work week. The observed number of defectives were 12, 4, 7, 14, and 10. So the sample mean is 9.4.

Thus a point estimate for μ, the hourly mean number of defectives, is 9.4.

8

8

Confidence Intervals

- An Interval Estimate states the range within which a population parameter "probably" lies.
- The interval within which a population parameter is expected to occur is called a Confidence Interval.
- The distance from the center of the confidence interval to the endpoint is called the **"Margin of Error"**
- The three confidence intervals that are used extensively are the 90%, 95% and 99%.

9

9

Confidence Intervals

- A 95% confidence interval means that about 95% of the similarly constructed intervals will contain the parameter being estimated, or 95% of the sample means for a specified sample size will lie within 1.96 standard deviations of the hypothesized population mean.
- For the 99% confidence interval, 99% of the sample means for a specified sample size will lie within 2.58 standard deviations of the hypothesized population mean.
- For the 90% confidence interval, 90% of the sample means for a specified sample size will lie within 1.645 standard deviations of the hypothesized population mean.

10

10

90%, 95% and 99% Confidence Intervals for µ

- The 90%, 95% and 99% confidence intervals for μ are constructed as follows when $n \geq 30$
- 90% CI for the population mean is given by

$$\overline{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

- 95% CI for the population mean is given by

$$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- 99% CI for the population mean is given by

$$\overline{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$$

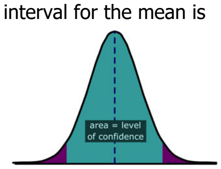
11

11

Constructing General Confidence Intervals for µ

- In general, a confidence interval for the mean is computed by:

$$\overline{X} \pm Z \frac{\sigma}{\sqrt{n}}$$



- This can also be thought of as:

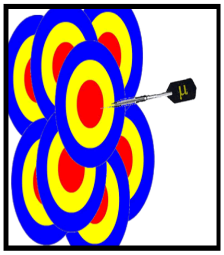
Point Estimator ± Margin of Error

12

12

The nature of Confidence Intervals

- The Population mean μ is fixed.
- The confidence interval is centered around the sample mean which is a Random Variable.
- So the Confidence Interval (Random Variable) is like a target trying hit a fixed dart (μ).



13

13

EXAMPLE

- The Dean wants to estimate the mean number of hours worked per week by students. A sample of 49 students showed a sample mean of 24 hours with a population standard deviation of 4 hours.
- The **point estimate** is 24 hours (sample mean).
- What is the **95% confidence interval** for the average number of hours worked per week by the students?

$$24 \pm 1.96\left(\frac{4}{\sqrt{49}}\right) = 24 \pm 1.12$$

22.88 ———+——— 25.12
24

14

14

EXAMPLE *continued*

- To interpret a confidence interval, use the following:
  - "We are <confidence level>% confident that <the parameter in context> is between <left endpoint> and <right endpoint>"
- We are 95% confident that the mean hours by students is between 22.88 and 25.12 hours per week.

15

15

EXAMPLE

- What is the **99% confidence interval** for the average number of hours worked per week by the students?

$$24 \pm 2.576\left(\frac{4}{\sqrt{49}}\right) = 24 \pm 1.47$$

22.53 ———+——— 25.47
24

- We are 99% confident that the mean hours by students is between 22.53 and 25.47 hours per week.
- Compare to 95% confidence interval. More confidence means larger margin of error (less precision)

16

16

Selecting a Sample Size

- There are 3 factors that determine the size of a sample, none of which has any direct relationship to the size of the population. They are:
  - The degree of confidence selected.
  - The maximum allowable error.
  - The variation of the population.

17

17

Sample Size for the Mean

- A convenient computational formula for determining n is:

$$n = \left(\frac{Z\sigma}{E}\right)^2$$

- where E is the allowable error (margin of error), Z is the z score associated with the degree of confidence selected, and σ is the sample deviation of the pilot survey.
- σ can be estimated by past data, target sample or range of data.

18

18

EXAMPLE

- The Dean wants to estimate with 99% confidence the mean number of hours worked per week by students with a margin of error (E) of 0.5 hours. Assume population standard deviation of 4 hours.

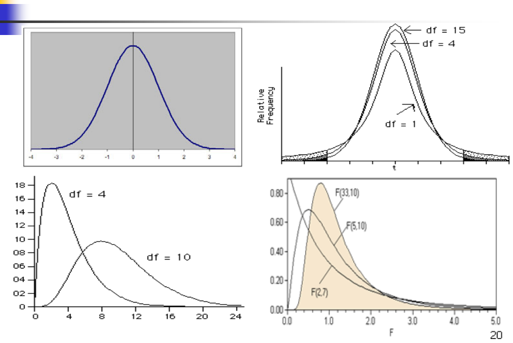
$$n = \left(\frac{(2.58)(24)}{0.5}\right) = 426.0096 = 427$$

- Note: round up to next whole number

19

19

Normal Family of Distributions: Z, t, χ^2, F



20

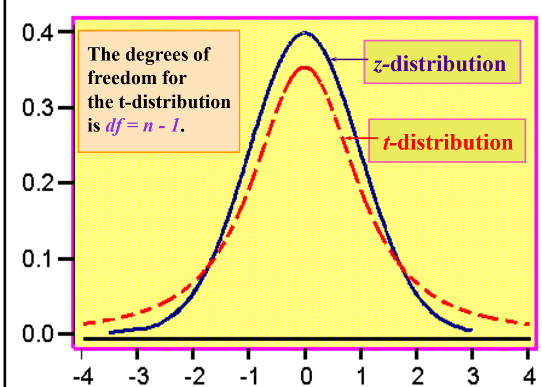
20

Characteristics of Student's *t*-Distribution

- The *t*-distribution has the following properties:
  - It is continuous, bell-shaped, and symmetrical about zero like the z-distribution.
  - There is a **family** of *t*-distributions sharing a mean of zero but having different standard deviations based on **degrees of freedom**.
  - The t-distribution is more spread out and flatter at the center than the z-distribution, but approaches the z-distribution as the sample size gets larger.

21

21



22

Confidence Interval for μ (σ unknown)

Formula to find a confidence interval using the t-distribution for the appropriate level of confidence:

$$\bar{X} \pm t\left(\frac{s}{\sqrt{n}}\right) \quad df = n - 1$$

23

23

Example – Confidence Interval

- In a random sample of 13 American adults, the mean waste recycled per person per day was 5.3 pounds and the standard deviation was 2.0 pounds.
- Assume the variable is normally distributed and construct a 95% confidence interval for μ.

24

24

Example- Confidence Interval

level of confidence = 95%
df=13-1=12
t=2.18

$$5.3 \pm 2.18 \frac{2.0}{\sqrt{13}}$$

$$5.3 \pm 1.2 = (4.1, 6.5)$$

25

25

Confidence Intervals, Population Proportions

- Point estimate for proportion of successes in population is: $\hat{p} = \frac{X}{n}$
- X is the number of successes in a sample of size n.
- Standard deviation of $\hat{p}$ is $\sqrt{\frac{(p)(1-p)}{n}}$
- Confidence Interval for p:

$$\hat{p} \pm Z \cdot \sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm Z \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

26

26

Population Proportion Example

- In a May 2006 AP/ISPOS Poll, 1000 adults were asked if "Over the next six months, do you expect that increases in the price of gasoline will cause financial hardship for you or your family, or not?"
- 700 of those sampled responded yes!
- Find the **sample proportion** and **margin of error** for this poll. (This means find a 95% confidence interval.)

27

27

Population Proportion Example

- Sample proportion

$$\hat{p} = \frac{700}{1000} = .70 = 70\%$$

- Margin of Error

$$MOE = 1.96\sqrt{\frac{.70(1-.70)}{1000}} = .028 = 2.8\%$$

28

28

Sample Size for the Proportion

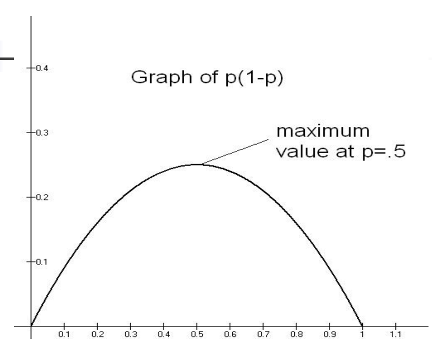
- A convenient computational formula for determining n is:

$$n = \left(p(1-p)\right)\left(\frac{Z}{E}\right)^2$$

- where E is the allowable margin of error, Z is the z-score associated with the degree of confidence selected, and p is the population proportion.
- If p is completely unknown, p can be set equal to ½ which maximizes the value of (p)(1-p) and guarantees the confidence interval will fall within the margin of error.

29

29



30

30

Example

- In polling, determine the minimum sample size needed to have a margin of error of 3% when p is unknown.

$$n = (.5)(1-.5)\left(\frac{1.96}{.03}\right)^2 = 1068$$

31

31

Example

- In polling, determine the minimum sample size needed to have a margin of error of 3% when p is known to be close to 1/4.

$$n = (.25)(1-.25)\left(\frac{1.96}{.03}\right)^2 = 801$$

32

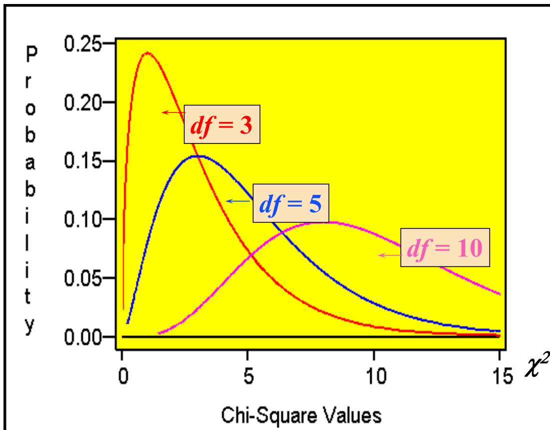
32

Characteristics of the Chi-Square Distribution

- The major characteristics of the chi-square distribution are:
  - It is positively skewed
  - It is non-negative
  - It is based on degrees of freedom
  - When the degrees of freedom change, a new distribution is created

33

33



34

Inference about Population Variance and Standard Deviation

- s^2 is an unbiased point estimator for σ^2
- s is a point estimator for σ
- Interval estimates and hypothesis testing for both σ^2 and σ require a new distribution – the χ^2 (Chi-square)

35

35

Distribution of s^2

- $\frac{(n-1)s^2}{\sigma^2}$ has a chi-square distribution
- n-1 is degrees of freedom
- s^2 is sample variance
- σ^2 is population variance

36

36

Confidence interval for σ^2

- Confidence is **NO**T symmetric since chi-square distribution is not symmetric. You must find separate left and right bounds.
- We can construct a confidence interval for σ^2

$$\left(\frac{(n-1)s^2}{\chi^2_R}, \frac{(n-1)s^2}{\chi^2_L} \right)$$

- Take square root of both endpoints to get confidence interval for σ, the population standard deviation.

37

37

Example

- In performance measurement of investments, standard deviation is a measure of volatility or risk.
- Twenty monthly returns from a mutual fund show an average monthly return of 1% and a sample standard deviation of 5%
- Find a 95% confidence interval for the monthly standard deviation of the mutual fund.

38

38

Example (cont)

- df = n-1 =19
- 95% CI for σ

$$\left(\sqrt{\frac{(19)5^2}{32.8523}}, \sqrt{\frac{(19)5^2}{8.90655}} \right) = (3.8, 7.3)$$

39

39

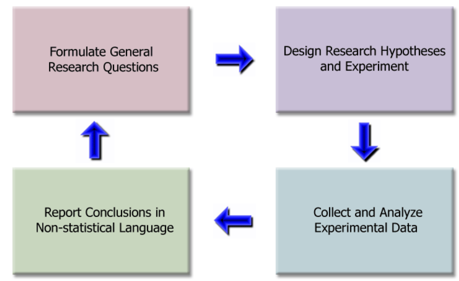
Inferential Statistics and Probability
a Holistic Approach

Chapter 9
One Population Hypothesis Testing

1

1

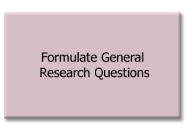
Procedures of Hypotheses Testing



2

2

Hypotheses Testing – Procedure 1



3

3

General Research Question

- Decide on a topic or phenomena that you want to research.
- Formulate general research questions based on the topic.
- Example:
  - Topic: Health Care Reform
  - Some General Questions:
    - Would a Single Payer Plan be less expensive than Private Insurance?
    - Do HMOs provide the same quality care as PPOs?
    - Would the public support mandated health coverage?

4

4

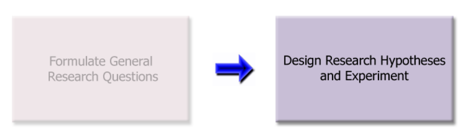
EXAMPLE – General Question

- A food company has a policy that the stated contents of a product match the actual results.
- A General Question might be "Does the stated net weight of a food product match (on average) the actual weight?"
- The quality control statistician could then decide to test various food products for accuracy.

5

5

Hypotheses Testing – Procedure 2



6

6

Hypothesis Testing Design

State Your Hypotheses
- Null Hypothesis
- Alternative Hypothesis

Determine Appropriate Model
- Test Statistic
- One or Two Tailed

Determine Decision Criteria
- α – Significance Level
- β and Power Analysis

7

7

What is a Hypothesis?

- Hypothesis: A statement about the value of a population parameter developed for the purpose of testing.
- Examples of hypotheses made about a population parameter are:
  - The mean monthly income for programmers is $9,000.
  - At least twenty percent of all juvenile offenders are caught and sentenced to prison.
  - The standard deviation for an investment portfolio is no more than 10 percent per month.

8

8

What is Hypothesis Testing?

- Hypothesis testing: A procedure, based on sample evidence and probability theory, used to determine whether the hypothesis is a reasonable statement and should not be rejected, or is unreasonable and should be rejected.

9

9

Hypothesis Testing Design

State Your Hypotheses
- Null Hypothesis
- Alternative Hypothesis

Determine Appropriate Model
- Test Statistic
- One or Two Tailed

Determine Decision Criteria
- α – Significance Level
- β and Power Analysis

10

10

Definitions

- Null Hypothesis H_0: A statement about the value of a population parameter that is assumed to be true for the purpose of testing.
- Alternative Hypothesis H_a: A statement about the value of a population parameter that is assumed to be true if the Null Hypothesis is rejected during testing.

11

11

Hypotheses written in words and population parameters

- Ho: The mean monthly income for programmers is $9,000.
- Ha: The mean monthly income for programmers is not $9,000.

$$Ho: \mu = 9000 \quad Ha: \mu \neq 9000$$

- Ho: At least 20% of all juvenile offenders sentenced to prison.
- Ha: Less than 20% of all juvenile offenders sentenced to prison.

$$Ho: p \geq 0.20 \quad Ha: p < 0.20$$

- Ho: The standard deviation for an investment portfolio is no more than 10 percent per month.
- Ha: The standard deviation for an investment portfolio is more than 10 percent per month.

$$Ho: \sigma \leq 10 \quad Ha: \sigma > 10$$

12

12

EXAMPLE – Stating Hypotheses

- A food company has a policy that the stated contents of a product match the actual results.
- The quality control statistician decides to test the claim that a 16 ounce bottle of Soy sauce contains on average 16 ounces.
- Ho: The mean amount of Soy Sauce is 16 ounces
- Ha: The mean amount of Soy Sauce is not 16 ounces.
- Ho: μ=16   Ha: $\mu \neq 16$

13

13

Hypothesis Testing Design

State Your Hypotheses
- Null Hypothesis | Alternative Hypothesis

Determine Appropriate Model
- Test Statistic | One or Two Tailed

Determine Decision Criteria
- α – Significance Level | β and Power Analysis

14

14

Definitions

- Statistical Model: A mathematical model that describes the behavior of the data being tested.
- Normal Family = the Standard Normal Distribution (Z) and functions of independent Standard Normal Distributions (eg: t, χ^2, F).
  - Most Statistical Models will be from the Normal Family due to the Central Limit Theorem.
- Model Assumptions: Criteria which must be satisfied to appropriately use a chosen Statistical Model.
- Test statistic: A value, determined from sample information, used to determine whether or not to reject the null hypothesis.

15

15

EXAMPLE – Choosing Model

- The quality control statistician decides to test the claim that a 16 ounce bottle of Soy sauce contains on average 16 ounces. We will assume the population standard is known
- Ho: μ=16   Ha: $\mu \neq 16$
- Model: One sample Z test of mean
- Test Statistic: $Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

16

16

Hypothesis Testing Design

State Your Hypotheses
- Null Hypothesis | Alternative Hypothesis

Determine Appropriate Model
- Test Statistic | One or Two Tailed

Determine Decision Criteria
- α – Significance Level | β and Power Analysis

17

17

Definitions

- Level of Significance: The probability of rejecting the null hypothesis when it is actually true. (signified by α)
- Type I Error: Rejecting the null hypothesis when it is actually true.
- Type II Error: Failing to reject the null hypothesis when it is actually false.

18

18

Outcomes of Hypothesis Testing

	Fail to Reject Ho	Reject Ho
Ho is true	Correct Decision	Type I error
Ho is False	Type II error	Correct Decision

19

19

EXAMPLE – Type I and Type II Errors

- Ho: The mean amount of Soy Sauce is 16 ounces
- Ha: The mean amount of Soy Sauce is not 16 ounces.
- Type I Error: The researcher **supports** the claim that the mean amount of soy sauce is not 16 ounces when the actual mean is 16 ounces. The company needlessly "fixes" a machine that is operating properly.
- Type II Error: The researcher **fails to support** the claim that the mean amount of soy sauce is not 16 ounces when the actual mean is not 16 ounces. The company fails to fix a machine that is not operating properly.

20

20

Hypothesis Testing Design

State Your Hypotheses

Null Hypothesis | Alternative Hypothesis

Determine Appropriate Model

Test Statistic | One or Two Tailed

Determine Decision Criteria

α – Significance Level | β and Power Analysis

21

21

Definitions

- Critical value(s): The dividing point(s) between the region where the null hypothesis is rejected and the region where it is not rejected. The critical value determines the decision rule.
- Rejection Region: Region(s) of the Statistical Model which contain the values of the Test Statistic where the Null Hypothesis will be rejected. The area of the Rejection Region = α

22

22

One-Tailed Tests of Significance

- A test is one-tailed when the alternate hypothesis, H_a , states a direction, such as:
  - H_0 : The mean income of females is less than or equal to the mean income of males.
  - H_a : The mean income of females is greater than males.
- Equality is part of H_0
- H_a determines which tail to test
  - H_a: $\mu>\mu_0$ means test upper tail.
  - H_a: $\mu<\mu_0$ means test lower tail.

23

23

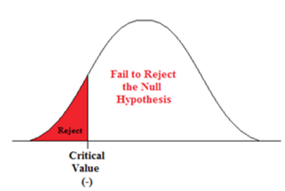
Left-tailed test

$$H_0 : \mu \geq \mu_0$$

$$H_a : \mu < \mu_0$$

$$\alpha = .05$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$



24

24

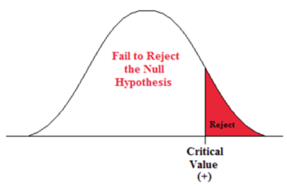
Right-tailed test

$$H_0 : \mu \le \mu_0$$

$$H_a : \mu > \mu_0$$

$$\alpha = .05$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$



25

25

Two-Tailed Tests of Significance

- A test is two-tailed when no direction is specified in the alternate hypothesis H_a , such as:
  - H_0 : The mean income of females is equal to the mean income of males.
  - H_a : The mean income of females is not equal to the mean income of the males.
- Equality is part of H_0
- H_a determines which tail to test
  - H_a: $\mu \neq \mu_0$ means test both tails.

26

26

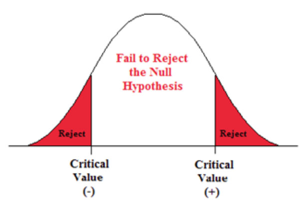
Two-tailed test

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

$$\alpha = .05 \quad \alpha/2 = .025$$

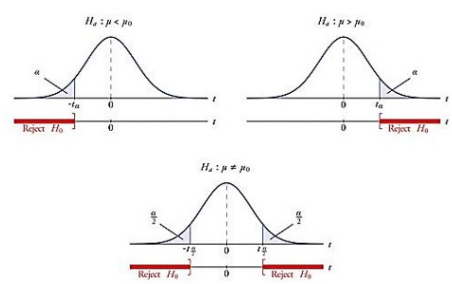
$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$



27

27

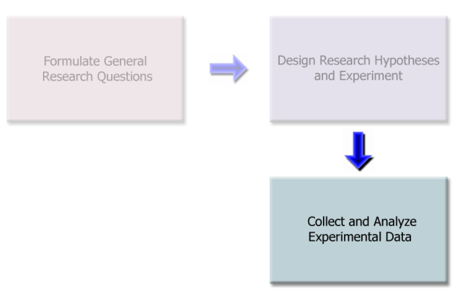
Left, Right and Two-Tailed Tests



28

28

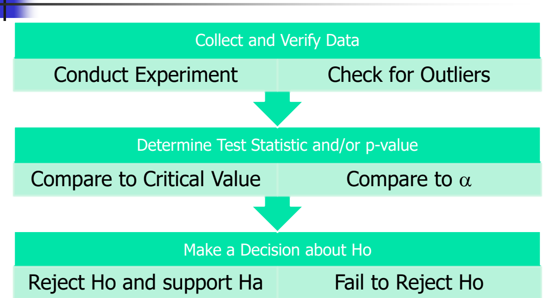
Hypotheses Testing – Procedure 3



29

29

Collect and Analyze Experimental Data



30

30

Collect and Analyze Experimental Data

Collect and Verify Data

Conduct Experiment | Check for Outliers

Determine Test Statistic and/or p-value

Compare to Critical Value | Compare to α

Make a Decision about Ho

Reject Ho and support Ha | Fail to Reject Ho

31

31

Outliers

- An outlier is data point that is far removed from the other entries in the data set.
- Outliers could be
  - Mistakes made in recording data
  - Data that don't belong in population
  - True rare events

32

32

Outliers have a dramatic effect on some statistics

- Example quarterly home sales for 10 realtors:

2 2 3 4 5 5 6 6 7 50

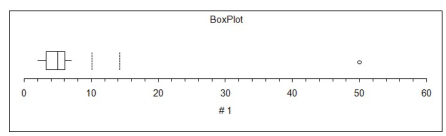
	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

33

33

Using Box Plot to find outliers

- The "box" is the region between the 1st and 3rd quartiles.
- Possible outliers are more than 1.5 IQR's from the box (inner fence)
- Probable outliers are more than 3 IQR's from the box (outer fence)
- In the box plot below, the dotted lines represent the "fences" that are 1.5 and 3 IQR's from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.



34

34

Using Z-score to detect outliers

- Calculate the mean and standard deviation without the suspected outlier.
- Calculate the Z-score of the suspected outlier.
- If the Z-score is more than 3 or less than -3, that data point is a probable outlier.

$$Z = \frac{50 - 4.4}{1.81} = 25.2$$

35

35

Outliers – what to do

- Remove or not remove, there is no clear answer.
- For some populations, outliers don't dramatically change the overall statistical analysis. Example: the tallest person in the world will not dramatically change the mean height of 10000 people.
- However, for some populations, a single outlier will have a dramatic effect on statistical analysis (called "**Black Swan**" by Nicholas Taleb) and inferential statistics may be invalid in analyzing these populations. Example: the richest person in the world will dramatically change the mean wealth of 10000 people.

36

36

Example – Analyze Data

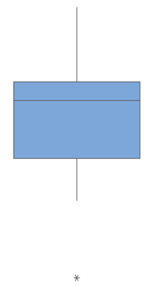
- In the Soy Sauce Example, a 36 bottles were measured, volume is in fluid ounces

- 14.51 15.16 15.28 15.33 15.36 15.42
- 15.43 15.45 15.49 15.59 15.60 15.61
- 15.62 15.63 15.71 15.81 15.87 16.00
- 16.01 16.02 16.05 16.06 16.06 16.09
- 16.09 16.11 16.16 16.16 16.27 16.31
- 16.35 16.36 16.45 16.72 16.75 16.79

37

37

Example – Analyze Data



- Although 14.51 might be a possible outlier and the data seems negatively skewed, the Central Limit Theorem assures that the sample mean will have a normal distribution

38

38

Collect and Analyze Experimental Data

Collect and Verify Data

Conduct Experiment | Check for Outliers

Determine Test Statistic and/or p-value

Compare to Critical Value | Compare to α

Make a Decision about Ho

Reject Ho and support Ha | Fail to Reject Ho

39

39

The logic of Hypothesis Testing

- This is a "Proof" by contradiction.
  - We assume H_o is true before observing data and design H_a to be the complement of H_o.
  - Observe the data (evidence). How unusual are these data under H_o?
  - If the data are too unusual, we have "proven" H_o is false: Reject H_o and go with H_a (Strong Statement)
  - If the data are not too unusual, we fail to reject H_o. This "proves" nothing and we say data are inconclusive. (Weak Statement)
  - We can never "prove" H_o , only "disprove" it.
  - "Prove" in statistics means support with the Alternative Hypothesis.
  - Note: It is **never correct** to say $(1-\alpha)$100% certain of our decision. (example: if α=.05, then we are not 95% certain if we Reject Ho.)

40

40

Test Statistic

- Test Statistic: A value calculated from the Data under the appropriate Statistical Model from the Data that can be compared to the Critical Value of the Hypothesis test
- If the Test Statistic fall in the Rejection Region, Ho is rejected.
- The Test Statistic will also be used to calculate the p-value as will be defined next.

41

41

Example - Testing for the Population Mean
Large Sample, Population Standard Deviation Known

- When testing for the population mean from a large sample and the population standard deviation is known, the test statistic is given by:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

42

42

p-value in Hypothesis Testing

- **p-Value**: the probability, assuming that the null hypothesis is true, of getting a value of the test statistic at least as extreme as the computed value for the test.
- If the p-value is smaller than the significance level, H_0 is rejected.
- If the p-value is larger than the significance level, H_0 is not rejected.

43

43

Comparing p-value to α

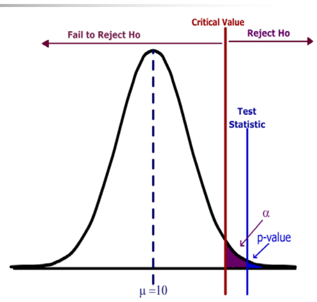
- Both **p-value** and α are probabilities.
- The **p-value** is determined by the **data**, and is the probability of getting results as extreme as the data assuming H_0 is true. Small values make one more likely to reject H_0.
- α is determined by **design**, and is the maximum probability the experimenter is willing to accept of rejecting a true H_0.
- Reject H_0 if p-value < α for ALL MODELS.

44

44

Graphic where decision is to Reject Ho

- Ho: μ = 10
  Ha: μ > 10
- Design: Critical Value is determined by significance level α.
- Data Analysis: p-value is determined by Test Statistic
- Test Statistic falls in Rejection Region.
- p-value (blue) < α (purple)
- Reject Ho.
- Strong statement: Data supports Alternative Hypothesis.

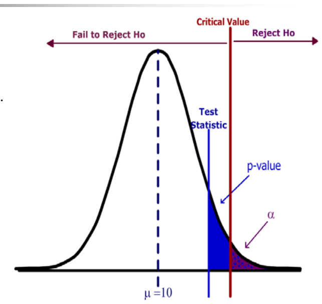


45

45

Graphic where decision is Fail to Reject Ho

- Ho: μ = 10
  Ha: μ > 10
- Design: Critical Value is determined by significance level α.
- Data Analysis: p-value is determined by Test Statistic
- Test Statistic falls in Non-rejection Region.
- p-value (blue) > α (purple)
- Fail to Reject Ho.
- Weak statement: Data is inconclusive and does not support Alternative Hypothesis.



46

46

EXAMPLE – General Question

- A food company has a policy that the stated contents of a product match the actual results.
- A General Question might be "Does the stated net weight of a food product match the actual weight?"
- The quality control statistician decides to test the 16 ounce bottle of Soy Sauce.

47

47

EXAMPLE – Design Experiment

- A sample of n=36 bottles will be selected hourly and the contents weighed. Assume σ = 0.5
- Ho: μ=16 Ha: $\mu \neq 16$
- The Statistical Model will be the one population test of mean using the Z Test Statistic.
- This model will be appropriate since the sample size insures the sample mean will have a Normal Distribution (Central Limit Theorem)
- We will choose a significance level of α = 5%

48

48

EXAMPLE – Conduct Experiment

- Last hour a sample of 36 bottles had a mean weight of 15.88 ounces.
- From past data, assume the population standard deviation is 0.5 ounces.
- Compute the Test Statistic

$$Z = [15.88 - 16]/[.5/\sqrt{36}] = -1.44$$

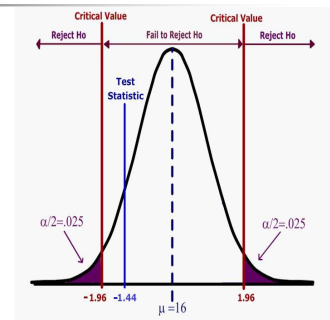
- For a two tailed test, The Critical Values are at Z = ±1.96

49

49

Decision – Critical Value Method

- This two-tailed test has two Critical Value and Two Rejection Regions
- The significance level (α) must be divided by 2 so that the sum of both purple areas is 0.05
- The Test Statistic does not fall in the Rejection Regions.
- Decision is **Fail to Reject Ho.**



50

50

Computation of the p-Value

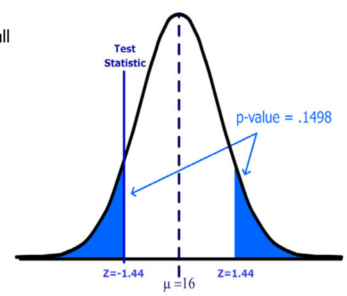
- One-Tailed Test: p-Value = P{z ≥ absolute value of the computed test statistic value}
- Two-Tailed Test: p-Value = 2P{z ≥ absolute value of the computed test statistic value}
- Example: Z= 1.44, and since it was a two-tailed test, then p-Value = 2P {z ≥ 1.44} = 0.0749) = .1498. Since .1498 > .05, do not reject H_0.

51

51

Decision – p-value Method

- The p-value for a two-tailed test must include all values (positive and negative) more extreme than the Test Statistic.
- p-value = .1498 which exceeds α = .05
- Decision is **Fail to Reject Ho.**



52

52

p-value form Minitab (shown as p)

One-Sample Z: weight

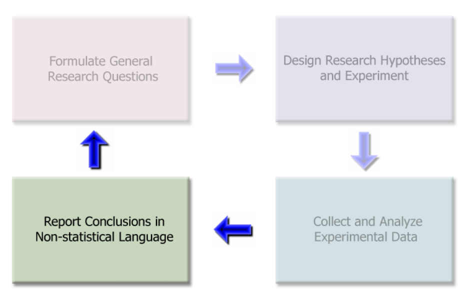
Test of μ = 16 vs ≠ 16
The assumed standard deviation = 0.5

Variable	N	Mean	StDev	SE Mean	Z	**P**
weight	36	15.8800	0.4877	0.0833	-1.44	**0.150**

53

53

Hypotheses Testing – Procedure 4



54

54

Converting Decision to Conclusion

- Conclusion if Decision is Reject Ho:
- <Ha in the Context of Problem>
- Conclusion if Decision is Fail to Reject Ho:
- "There is insufficient evidence to conclude" <Ha in the Context of Problem>

55

55

Example - Conclusion

- Decision: Fail to Reject Ho
- There is insufficient evidence to conclude that the mean amount of soy sauce being filled into bottles is not 16 ounces.
- There is insufficient evidence to conclude machine that fills 16 ounce soy sauce bottles is operating improperly.

56

56

Conclusions

- Conclusions need to
  - Be consistent with the results of the Hypothesis Test.
  - Use language that is clearly understood in the context of the problem.
  - Limit the inference to the population that was sampled.
  - Report sampling methods that could question the integrity of the random sample assumption.
  - Conclusions should address the potential or necessity of further research, sending the process back to the first procedure.

57

57

Conclusions need to be consistent with the results of the Hypothesis Test.

- Rejecting Ho requires a strong statement in support of Ha.
- Failing to Reject Ho does NOT support Ho, but requires a weak statement of insufficient evidence to support Ha.
- Example:
  - The researcher wants to support the claim that, on average, students send more than 1000 text messages per month
  - Ho: $\mu=1000$ Ha: $\mu>1000$
  - Conclusion if Ho is rejected: The mean number of text messages sent by students exceeds 1000.
  - Conclusion if Ho is not rejected: There is insufficient evidence to support the claim that the mean number of text messages sent by students exceeds 1000.

58

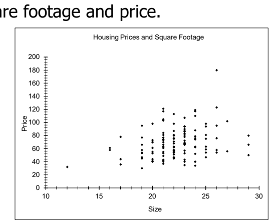
58

Conclusions need to use language that is clearly understood in the context of the problem.

- Avoid technical or statistical language.
- Refer to the language of the original general question.
- Compare these two conclusions from a test of correlation between home prices square footage and price.

Conclusion 1: By rejecting the Null Hypothesis we are inferring that the Alterative Hypothesis is supported and that there exists a significant correlation between the independent and dependent variables in the original problem comparing home prices to square footage.

Conclusion 2: Homes with more square footage generally have higher prices.



59

59

Conclusions need to limit the inference to the population that was sampled.

- If a survey was taken of a sub-group of population, then the inference applies to the subgroup.
- Example
  - Studies by pharmaceutical companies will only test adult patients, making it difficult to determine effective dosage and side effects for children.
  - "In the absence of data, doctors use their medical judgment to decide on a particular drug and dose for children. 'Some doctors stay away from drugs, which could deny needed treatment,' Blumer says. "Generally, we take our best guess based on what's been done before."
  - "The antibiotic chloramphenicol was widely used in adults to treat infections resistant to penicillin. But many newborn babies died after receiving the drug because their immature livers couldn't break down the antibiotic."

source: FDA Consumer Magazine – Jan/Feb 2003

60

60

Conclusions need to report sampling methods that could question the integrity of the random sample assumption.

- Be aware of how the sample was obtained. Here are some examples of pitfalls:
  - Telephone polling was found to under-sample young people during the 2008 presidential campaign because of the increase in cell phone only households. Since young people were more likely to favor Obama, this caused bias in the polling numbers.
  - Sampling that didn't occur over the weekend may exclude many full time workers.
  - Self-selected and unverified polls (like ratemyprofessors.com) could contain immeasurable bias.

61

61

Conclusions should address the potential or necessity of further research, sending the process back to the first procedure.

- Answers often lead to new questions.
- If changes are recommended in a researcher's conclusion, then further research is usually needed to analyze the impact and effectiveness of the implemented changes.
- There may have been limitations in the original research project (such as funding resources, sampling techniques, unavailability of data) that warrants more a comprehensive study.
  - Example: A math department modifies is curriculum based on a performance statistics for an experimental course. The department would want to do further study of student outcomes to assess the effectiveness of the new program.

62

62

Soy Sauce Example - Conclusion

- There is insufficient evidence to conclude that the machine that fills 16 ounce soy sauce bottles is operating improperly.
- This conclusion is based on 36 measurements taken during a single hour's production run.
- We recommend continued monitoring of the machine during different employee shifts to account for the possibility of potential human error.

63

63

Procedures of Hypotheses Testing

Formulate General Research Questions → Design Research Hypotheses and Experiment → Collect and Analyze Experimental Data → Report Conclusions in Non-statistical Language → Formulate General Research Questions

64

64

Hypothesis Testing Design

State Your Hypotheses	
Null Hypothesis	Alternative Hypothesis

Determine Appropriate Model	
Test Statistic	One or Two Tailed

Determine Decision Criteria	
α – Significance Level	β and Power Analysis

65

65

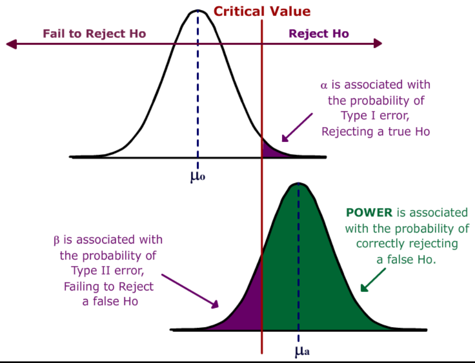
Statistical Power and Type II error

	Fail to Reject Ho	Reject Ho
Ho is true	$1-\alpha$	α Type I error
Ho is False	β Type II error	$1-\beta$ Power

66

66

Graph of "Four Outcomes"



67

67

Statistical Power (continued)

- Power is the probability of rejecting a false Ho, when $\mu = \mu_a$
- Power depends on:
  - Effect size $|\mu_o - \mu_a|$
  - Choice of α
  - Sample size
  - Standard deviation
  - Choice of statistical test

68

68

Statistical Power Example

- Bus brake pads are claimed to last on average at least 60,000 miles and the company wants to test this claim.
- The bus company considers a "practical" value for purposes of bus safety to be that the pads at least 58,000 miles.
- If the standard deviation is 5,000 and the sample size is 50, find the Power of the test when the mean is really 58,000 miles. Assume α = .05

69

69

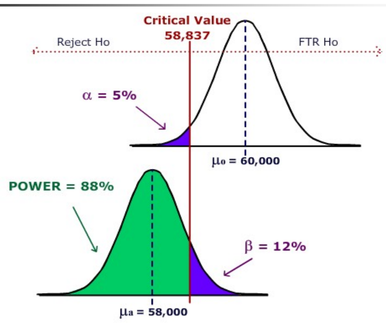
Statistical Power Example

- Set up the test
  - Ho: μ >= 60,000 miles
  - Ha: μ < 60,000 miles
  - α = 5%
- Determine the Critical Value
  - Reject Ho if $\overline{X} > 58{,}837$
- Calculate β and Power
  - β = 12%
  - Power = 1 – β = 88%

70

70

Statistical Power Example



71

71

New Models, Similar Procedures

- The **procedures** outlined for the One Sample Z test of Mean (with known population standard deviation) will apply to other models as well.
- Examples of some other one population models:
  - One Sample t-test of mean, population standard deviation unknown.
  - One sample Z-test of proportion.
  - One sample Chi-square test of variance (or standard deviation)

72

72

Testing for the Population Mean: Population Standard Deviation Unknown

- Model: One sample t-test of mean
- $$t = \frac{\bar{X} - \mu_o}{s/\sqrt{n}}$$
- The degrees of freedom for the test is n-1.
- Assumptions: $\bar{X}$ has a Normal Distribution

73

73

Decision Rules

- Like the normal distribution, the logic for one and two tail testing is the same.
- For a two-tail test using the *t*-distribution, you will reject the null hypothesis when the value of the test statistic is greater than $t_{df,\alpha/2}$ or if it is less than - $t_{df,\alpha/2}$
- For a left-tail test using the *t*-distribution, you will reject the null hypothesis when the value of the test statistic is less than $-t_{df,\alpha}$
- For a right-tail test using the *t*-distribution, you will reject the null hypothesis when the value of the test statistic is greater than $t_{df,\alpha}$

74

74

Example – one population test of mean, σ unknown

- Humerus bones from the same species have approximately the same length-to-width ratios. When fossils of humerus bones are discovered, archaeologists can determine the species by examining this ratio. It is known that Species A has a mean ratio of 9.6. A similar Species B has a mean ratio of 9.1 and is often confused with Species A.
- 21 humerus bones were unearthed in an area that was originally thought to be inhabited Species A. (Assume all unearthed bones are from the same species.)
- Design a hypothesis test where the alternative claim would be the humerus bones were not from Species A.
- Determine the power of this test if the bones actually came from Species B (assume a standard deviation of 0.7)
- Conduct the test using at a 5% significance level and state overall conclusions.

75

75

Example – Designing Test

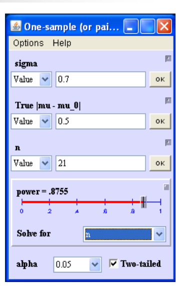
- Research Hypotheses
  - Ho: The humerus bones are from Species A
  - Ha: The humerus bones are not from Species A
- In terms of the population mean
  - Ho: $\mu = 9.6$
  - Ha: $\mu \neq 9.6$
- Significance level
  - $\alpha = .05$
- Test Statistic (Model)
  - One sample t-test of mean

76

76

Example - Power Analysis

- Information needed for Power Calculation
  - $\mu_o = 9.6$ (Species A)
  - $\mu_a = 9.1$ (Species B)
  - Effect Size = $|\mu_o - \mu_a| = 0.5$
  - $\sigma = 0.7$ (given)
  - $\alpha = .05$
  - n = 21 (sample size)
  - Two tailed test
- Results using online Power Calculator*
  - Power =.8755
  - β = 1 - Power = .1245
  - If humerus bones are from Species B, test has an 87.55% chance of correctly rejecting Ho and a maximum Type II error of 12.45%

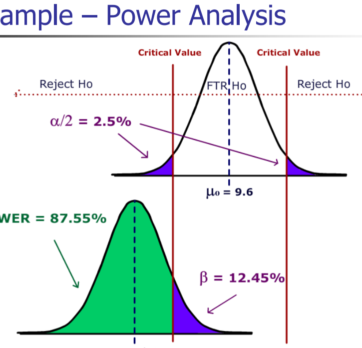


*source: Russ Lenth, University of Iowa – http://www.stat.uiowa.edu/~rlenth/Power/

77

77

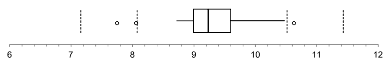
Example – Power Analysis



78

78

Example – Output of Data Analysis



Hypothesis Test: Mean vs. Hypothesized Value

9.60000	hypothesized value
9.26190	mean Data
0.66700	std. dev.
0.14555	std. error
21	n
20	df
-2.32	t
.0308	p-value (two-tailed)

P-value = .0308
α =.05
Since p-value < α
Ho is rejected and we support Ha.

79

79

Example - Conclusions

- Results:
  - The evidence supports the claim (pvalue<.05) that the humerus bones are not from Species A.
- Sampling Methodology:
  - We are assuming since the bones were unearthed in the same location, they came from the same species.
- Limitations:
  - A small sample size limited the power of the test, which prevented us from making a more definitive conclusion.
- Further Research
  - Test if the bone are from Species B or another unknown species.
  - Test to see if bones are the same age to support the sampling methodology.

80

80

Tests Concerning Proportion

- Proportion: A fraction or percentage that indicates the part of the population or sample having a particular trait of interest.
- The population proportion is denoted by p.
- The sample proportion is denoted by $\hat{p}$ where

$$\hat{p} = \frac{\text{number of successes in the sample}}{\text{number sampled}}$$

81

81

Test Statistic for Testing a Single Population Proportion

- If sample size is sufficiently large, $\hat{p}$ has an approximately normal distribution. This approximation is reasonable if $np_o \geq 10$ and $n(1-p_o) \geq 10$

$$Z = \frac{\hat{p} - p_o}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

p_o = *population proportion under Ho*

$\hat{p}$ = *sample proportion*

82

82

Example

- In the past, 15% of the mail order solicitations for a certain charity resulted in a financial contribution.
- A new solicitation letter has been drafted and will be sent to a random sample of potential donors.
- A hypothesis test will be run to determine if the new letter is more effective.
- Determine the sample size so that:
  - The test can be run at the 5% significance level.
  - If the letter has an 18% success rate, (an effect size of 3%), the power of the test will be 95%
- After determining the sample size, conduct the test.

83

83

Example – Designing Test

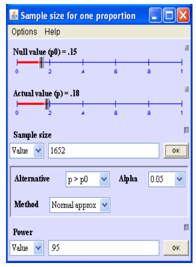
- Research Hypotheses
  - Ho: The new letter is not more effective.
  - Ha: The new letter is more effective.
- In terms of the population proportion
  - Ho: p = 0.15
  - Ha: p > 0.15
- Significance level
  - α =.05
- Test Statistic (Model)
  - One sample Z-test of proportion

84

84

Example - Power Analysis

- Information needed for Sample Size Calculation
  - p_o = 0.15 (current letter)
  - p_a = 0.18 (potential new letter)
  - Effect Size =| p_o - p_a | = 0.03
  - Desired Power = 0.95
  - α = .05
  - One tailed test



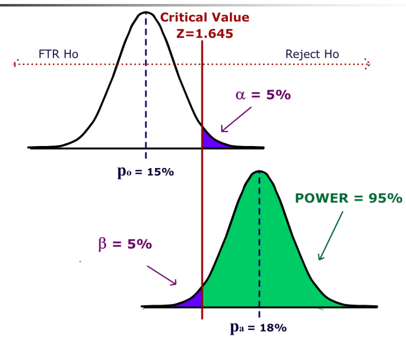
- Results using online Power Calculator*
  - Sample size = 1652
  - The charity should send out 1652 new solicitation letters to potential donors and run the test.

*source: Russ Lenth, University of Iowa – http://www.stat.uiowa.edu/~rlenth/Power/

85

85

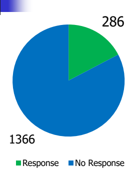
Example – Power Analysis



86

86

Example – Output of Data Analysis



Hypothesis test for proportion vs hypothesized value

Observed	Hypothesized	
0.1731	0.15	p (as decimal)
286/1652	248/1652	p (as fraction)
286.	247.8	X
1652	1652	n
	0.0088	std. error
	2.63	z
	.0042	p-value (one-tailed, upper)

- P-value = .0042
- α =0.05
- Since p-value < α, Ho is rejected and we support Ha.

87

87

EXAMPLE
Critical Value Alternative Method

- Critical Value =1.645 (95th percentile of the Normal Distribution.)
- H_0 is rejected if Z > 1.645
- Test Statistic: $$Z = \frac{\left(\frac{286}{1652} - .15\right)}{\sqrt{\frac{(.15)(.85)}{1652}}} = 2.63$$
- Since Z = 2.63 > 1.645, H_0 is rejected. The new letter is more effective.

88

88

Example - Conclusions

- Results:
  - The evidence supports the claim (pvalue<.01) that the new letter is more effective.
- Sampling Methodology:
  - The 1652 test letters were selected as a random sample from the charity's mailing list. All letters were sent at the same time period.
- Limitations:
  - The letters needed to be sent in a specific time period, so we were not able to control for seasonal or economic factors.
- Further Research
  - Test both solicitation methods over the entire year to eliminate seasonal effects.
  - Send the old letter to another random sample to create a control group.

89

89

Test for Variance or Standard Deviation vs. Hypothesized Value

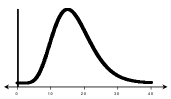
- We often want to make a claim about the variability, volatility or consistency of a population random variable.
- Hypothesized values for population variance σ^2 or standard deviation σ are tested with the χ^2 distribution.
- Examples of Hypotheses:
  - Ho: σ = 10 Ha: $\sigma \neq 10$
  - Ho: σ^2 = 100 Ha: $\sigma^2 > 100$
- The sample variance s^2 is used in calculating the Test Statistic.

90

90

Test Statistic uses χ^2 distribution

s^2 is the test statistic for the population variance. Its sampling distribution is a χ^2 distribution with n-1 d.f.



$$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2}$$

91

91

Example

- A state school administrator claims that the standard deviation of test scores for 8th grade students who took a life-science assessment test is less than 30, meaning the results for the class show consistency.
- An auditor wants to support that claim by analyzing 41 students recent test scores, shown here:

57	75	86	92	101	108	110	120	155
63	77	88	96	102	108	111	122	
66	78	88	96	107	109	115	135	
68	81	92	98	107	109	115	137	
72	82	92	99	107	110	118	139	

- The test will be run at 1% significance level.

92

92

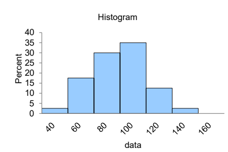
Example – Designing Test

- Research Hypotheses
  - Ho: Standard deviation for test scores equals 30.
  - Ha: Standard deviation for test scores is less than 30.
- In terms of the population variance
  - Ho: $\sigma^2 = 900$
  - Ha: $\sigma^2 < 900$
- Significance level
  - $\alpha = .01$
- Test Statistic (Model)
  - One sample Chi-square test of variance

93

93

Example – Output of Data Analysis



Chi-square Variance Test

900.000 hypothesized variance
469.426 observed variance of Data
41 n
40 df
20.86 chi-square

.0054 p-value (one-tailed, lower)

- p-value = .0054
- α =0.01
- Since p-value < α, Ho is rejected and we support Ha.

94

94

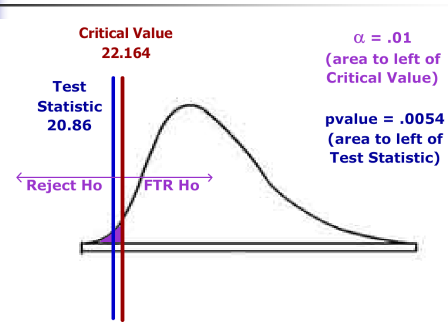
EXAMPLE
Critical Value Alternative Method

- Critical Value =22.164 (1st percentile of the Chi-square Distribution.)
- H_0 is rejected if $\chi^2 < 22.164$
- Test Statistic: $\chi^2 = \frac{(40)(469.426)}{900} = 20.86$
- Since Z = 20.86< 22.164, H_0 is rejected. The claim that the standard deviation is under 30 is supported.

95

95

Example – Decision Graph



96

96

Example - Conclusions

- Results:
  - The evidence supports the claim (pvalue<.01) that the standard deviation for 8th grade test scores is less than 30.
- Sampling Methodology:
  - The 41 test scores were the results of the recently administered exam to the 8th grade students.
- Limitations:
  - Since the exams were for the current class only, there is no assurance that future classes will achieve similar results.
- Further Research
  - Compare results to other schools that administered the same exam.
  - Continue to analyze future class exams to see if the claim is holding true.

97

97

Inferential Statistics and Probability
a Holistic Approach

Chapter 10

Two Population Inference

1

1

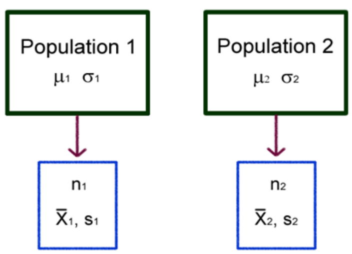
Comparing two population means

- Four models
  - Independent Sampling
    - Known population variances
      - **Two sample Z - test**
    - The 2 population variances are equal
      - **Pooled variance t-test**
    - The 2 population variances are unequal
      - **t-test for unequal variances**
  - Dependent Sampling
      - **Matched Pairs t-test**

2

2

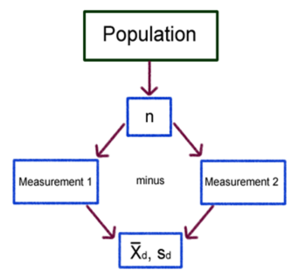
Independent Sampling



3

3

Dependent sampling



4

4

Difference of Two Population means

- $\overline{X}_1 - \overline{X}_2$ is Random Variable
- $\overline{X}_1 - \overline{X}_2$ is a point estimator for $\mu 1 - \mu 2$
- The standard deviation is given by the formula $\sqrt{\frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2}}$
- If n1 and n2 are sufficiently large, $\overline{X}_1 - \overline{X}_2$ follows a normal distribution.

5

5

Difference between two means – known population variances

- If both σ_1 and σ_2 are known and the two populations are independently selected, this test can be run.
- Test Statistic:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

6

6

Example 1

- Are larger houses more likely to have pools?
- The housing data square footage (size) was split into two groups by pool (Y/N).
- Can you support the claim that homes with pools have more square feet than the homes without pools. Let α = .01

7

7

EXAMPLE 1 – 2 sample Z-test of mean

$$H_o: \mu_1 \le \mu_2 \quad H_a: \mu_1 > \mu_2$$
$$H_o: \mu_1 - \mu_2 \le 0 \quad H_a: \mu_1 - \mu_2 > 0$$

$$\alpha = .01 \qquad Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

H_0 is rejected if *Z>2.326* *(or pvalue* $< \alpha$)

8

8

EXAMPLE 1 *Data*

- Population 1
  Size with pool
- Sample size = 130
- Sample mean = 26.25
- Pop Std Dev = 6.93

- Population 2
  Size without pool
- Sample size = 95
- Sample mean = 23.04
- Pop Std Dev = 4.55

9

9

EXAMPLE 1 *DATA*

$$Z = \frac{(26.25 - 23.04) - 0}{\sqrt{\frac{6.93^2}{130} + \frac{4.55^2}{95}}} = 4.19$$

- p-value = P(Z>4.19) = 0.0000137 < 0.01
- Decision: Reject Ho
- Conclusion: Homes with pools have more mean square footage.

10

10

EXAMPLE 1 *p-value method*

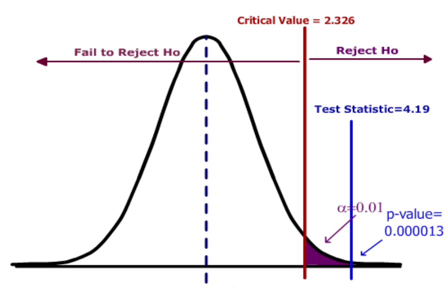
- Using Technology Reject Ho if the p-value < α

	Sq ft with pool	*Sq ft no pool*
Mean	26.25	23.04
Std Dev	6.93	4.55
Observations	130	95
Hypothesized Mean Difference	0	
Z	4.19	
p-value	**0.0000137**	

11

11

EXAMPLE 1 – Results/Decision



12

12

Pooled variance t-test

- To conduct this test, three assumptions are required:
  - The populations must be normally or approximately normally distributed (or central limit theorem must apply).
  - The sampling of populations must be **independent.**
  - The **population variances** must be **equal**.

13

13

Pooled Sample Variance and Test Statistic

- Pooled Sample Variance: $$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$
- Test Statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

14

14

EXAMPLE 2

- A recent EPA study compared the highway fuel economy of domestic and imported passenger cars.
- A sample of 12 imported cars revealed a mean of 35.76 mpg with a standard deviation of 3.86.
- A sample of 15 domestic cars revealed a mean of 33.59 mpg with a standard deviation of 2.16 mpg.
- At the .05 significance level can the EPA conclude that the mpg is higher on the imported cars? *(Let subscript 2 be associated with domestic cars.)*

15

15

EXAMPLE 2 – critical value method

- : $H_o: \mu_1 \le \mu_2 \quad H_a: \mu_1 > \mu_2$
- : α=.05
- : $t = (\bar{X}_1 - \bar{X}_2)/(s_p\sqrt{1/n_1 + 1/n_2})$
- : H_0 is rejected if t>1.708, df=25
- : t=1.85 H_0 is rejected. Imports have a higher mean mpg than domestic cars.

16

16

t-test when variances are not equal.

- Test statistic: $$t' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$
- Degrees of freedom: $$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}\right]}$$
- This test (also known as the Welch-Aspin Test) has **less power** then the prior test and should only be used when it is clear the population variances are different.

17

17

EXAMPLE 2

- : $H_o: \mu_1 \le \mu_2 \quad H_a: \mu_1 > \mu_2$
- : α=.05
- : t' test
- : H_0 is rejected if t>1.746, df=16
- : t'=1.74 H_0 is not rejected. There is insufficient sample evidence to claim a higher mpg on the imported cars.

18

18

Using Technology

- Decision Rule: Reject H_o if pvalue<α
- Megastat: Compare Two Independent Groups
- Use Equal Variance or Unequal Variance Test
- Use Original Data or Summarized Data

domestic 29.8 33.3 34.7 37.4 34.4 32.7 30.2 36.2 35.5 34.6 33.2 35.1 33.6 31.3 31.9

import 39.0 35.1 39.1 32.2 35.6 35.5 40.8 34.7 33.2 29.4 42.3 32.2

19

19

Pooled Variance t-test

- Minitab output
- p-value = 0.038
- p-value < α = .05
- Reject Ho

```
Two-sample T for MPG

TYPE      N   Mean  StDev  SE Mean
import   12  35.76   3.86      1.1
US car   15  33.59   2.16     0.56

Difference = µ (import) - µ (US car)
Estimate for difference:  2.16
95% lower bound for difference:  0.16
T-Test of difference = 0 (vs >): T-Value = 1.85
P-Value = 0.038  DF = 25
Both use Pooled StDev = 3.0264
```

20

20

Unequal Variances t-test

- Minitab output
- p-value = 0.051
- p-value < α = .05
- Fail to Reject Ho

```
Two-sample T for MPG

TYPE      N   Mean  StDev  SE Mean
import   12  35.76   3.86      1.1
US car   15  33.59   2.16     0.56

Difference = µ (import) - µ (US car)
Estimate for difference:  2.16
95% lower bound for difference:  -0.01
T-Test of difference = 0 (vs >): T-Value = 1.74
P-Value = 0.051  DF = 16
```

21

21

Hypothesis Testing - Paired Observations

- Independent samples are samples that are not related in any way.
- Dependent samples are samples that are paired or related in some fashion.
  - For example, if you wished to buy a car you would look at the *same* car at two (or more) *different* dealerships and compare the prices.
- Use the following test when the samples are dependent:

22

22

Hypothesis Testing Involving Paired Observations

$$t = \frac{\overline{X}_d - \mu_d}{s_d / \sqrt{n}}$$

- where $\overline{X}_d$ is the average of the differences
- s_d is the standard deviation of the differences
- n is the number of pairs (differences)

23

23

EXAMPLE 3

- An independent testing agency is comparing the daily rental cost for renting a compact car from Hertz and Avis.
- A random sample of 15 cities is obtained and the following rental information obtained.
- At the .05 significance level can the testing agency conclude that there is a difference in the rental charged?

24

24

Example 3 – continued

- Data for Hertz

$\overline{X}_1 = 46.67$

$s_1 = 5.23$

- Data for Avis

$\overline{X}_2 = 44.87$

$s_2 = 5.62$

City	Hertz	Avis
Atlanta	42	40
Baltimore	51	47
Boston	46	42
Chicago	56	52
Cleveland	45	43
Denver	48	48
Dallas	56	54
Honolulu	37	32
Los Angeles	51	48
Kansas City	45	48
Miami	41	39
New York	44	42
San Francisco	48	45
Seattle	46	50
Washington DC	44	43

25

25

Example 3 - continued

By taking the difference of each pair, variability (measured by standard deviation) is reduced.

$\overline{X}_d = 1.80$

$s_d = 2.513$

$n = 15$

City	Hertz	Avis	Difference
Atlanta	42	40	**2**
Baltimore	51	47	**4**
Boston	46	42	**4**
Chicago	56	52	**4**
Cleveland	45	43	**2**
Denver	48	48	**0**
Dallas	56	54	**2**
Honolulu	37	32	**5**
Los Angeles	51	48	**3**
Kansas City	45	48	**-3**
Miami	41	39	**2**
New York	44	42	**2**
San Francisco	48	45	**3**
Seattle	46	50	**-4**
Washington DC	44	43	**1**

26

26

EXAMPLE 3 *continued*

- $H_0: \mu_d = 0 \qquad H_1: \mu_d \neq 0$
- α=.05
- Matched pairs t test, df=14
- H_0 is rejected if t<-2.145 or t>2.145
- $t = (1.80)/[2.513/\sqrt{15}] = 2.77$
- Reject H_0.
- There is a difference in mean price for compact cars between Hertz and Avis. Avis has lower mean prices.

27

27

Megastat Output – Example 3

Hypothesis Test: Paired Observations

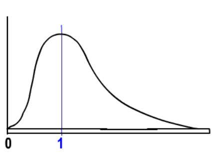
0.000	hypothesized value
46.667	mean Hertz
44.867	mean Avis
1.800	mean difference (Hertz - Avis)
2.513	std. dev.
0.649	std. error
15	n
14	df
2.77	t
.0149	p-value (two-tailed)

28

28

Characteristics of F-Distribution

- There is a "family" of *F* Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- *F* cannot be negative, and it is a continuous distribution.
- The *F* distribution is positively skewed.
- Its values range from 0 *to* ∞. As $F \to \infty$ the curve approaches the X-axis.



29

29

Test for Equal Variances

- For the two tail test, the test statistic is given by: $$F = \frac{S_i^2}{S_j^2}$$
- s_i^2 *and* s_j^2 are the sample variances for the two populations.
- There are 2 sets of degrees of freedom: n_i-1 for the numerator, n_j-1 for the denominator

30

30

EXAMPLE 4

- A stockbroker at brokerage firm, reported that the mean rate of return on a sample of 10 software stocks was 12.6 percent with a standard deviation of 4.9 percent.
- The mean rate of return on a sample of 8 utility stocks was 10.9 percent with a standard deviation of 3.5 percent.
- At the .05 significance level, can the broker conclude that there is more variation in the software stocks?

31

31

Test Statistic depends on Hypotheses

Hypotheses	Test Statistic
$H_o: \sigma_1 \geq \sigma_2$ $H_a: \sigma_1 < \sigma_2$	$F = \frac{s_2^2}{s_1^2}$ $\quad use\ \alpha\ table$
$H_o: \sigma_1 \leq \sigma_2$ $H_a: \sigma_1 > \sigma_2$	$F = \frac{s_1^2}{s_2^2}$ $\quad use\ \alpha\ table$
$H_o: \sigma_1 = \sigma_2$ $H_a: \sigma_1 \neq \sigma_2$	$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$ $\quad use\ \alpha/2\ table$

32

32

EXAMPLE 4 *continued*

- : $H_o: \sigma_1 \leq \sigma_2 \quad H_a: \sigma_1 > \sigma_2$
- : α =.05
- : F-test
- : H_0 is rejected if F>3.68, df=(9,7)
- : F=4.9²/3.5² =1.96 → Fail to Reject H_0.
- There is insufficient evidence to claim more variation in the software stock.

33

33

Minitab Example

- Using Minitab– Test for equal variances under two population independent samples test and click the box to test for equality of variances
- The default p-value is a two-tailed test, so take one-half reported p-value for one-tailed tests
- Example – Domestic vs Import Data
- $H_o: \sigma_1 = \sigma_2 \quad H_a: \sigma_1 \neq \sigma_2$
- α =0.05
- Reject Ho means use unequal variance t-test
- FTR Ho means use pooled variance t-test

34

34

Minitab Output

Test

Null hypothesis $\quad H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis $\quad H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level $\quad \alpha = 0.05$

Method Test

	Statistic	DF1	DF2	P-Value
F	3.20	11	14	0.044

pvalue <0.05
Reject Ho

Use unequal variance t-test to compare means.

35

35

Comparing two proportions

- Suppose we take a sample of n_1 from population 1 and n_2 from population 2.
- Let X_1 be the number of success in sample 1 and X_2 be the number of success in sample 2.
- The sample proportions are then calculated for each group.

$$\hat{p}_1 = \frac{X_1}{n_1} \qquad \hat{p}_2 = \frac{X_2}{n_2}$$

36

36

Hypothesis testing for 2 Proportions

- In conducting a Hypothesis test where the Null hypothesis assumes equal proportions, it is best practice to pool or combine the sample proportions into a single estimated proportion, and use an estimated standard error.

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} \qquad s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}$$

37

37

Hypothesis testing for 2 Proportions

- The test statistic will have a Normal Distribution as long as there are at least 10 successes and 10 failures in both samples.

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}}$$

38

38

Example

- In an August 2016 Study, Pew Research asked the sampled Americans if background checks required at gun stores should be made universal extended to all sales of guns between private owners or at gun shows.
- 772 out 990 men said yes, while 857 out of 1020 women said yes.
- Is there a difference in the proportion of men and women who support universal background checks for purchasing guns? Design and conduct the test with a significance level of 1%.

39

39

Example (Design)

- **Ho: $p_m = p_w$** (There is no difference in the proportion of support for background checks by gender)
  Ha: $p_m \neq p_w$ (There is a difference in the proportion of support for background checks by gender)
- **Model:**
  Two proportion Z test. This is a two-tailed test with a = 0.01.
- **Model Assumptions:** for men there are 772 yes and 218 no. For women there are 857 yes and 16 no. Since all these numbers exceed 10, the model is appropriate.
- **Decision Rules:**
  Critical Value Method - Reject Ho if Z > 2.58 or Z < -2.58.
  P-value method - Reject Ho if p-value <0.01

40

40

Example (Results)

$$\hat{p}_m = \frac{772}{990} = 0.780 \qquad \hat{p}_w = \frac{857}{1020} = 0.840$$

$$\bar{p} = \frac{772+857}{990+1020} = 0.810 \qquad Z = \frac{(0.780-0.840)-0}{\sqrt{\frac{0.810(1-0.810)}{990} + \frac{0.810(1-0.810)}{1020}}} = -3.45$$

p-value = 0.0005 < α
Reject Ho Under both methods.

Conclusion: There is a difference in the proportion of support for background checks by gender. Women are more likely to support background checks.

41

41

Inferential Statistics and Probability
a Holistic Approach

Chapter 11
Chi-square Tests for Categorical Data

1

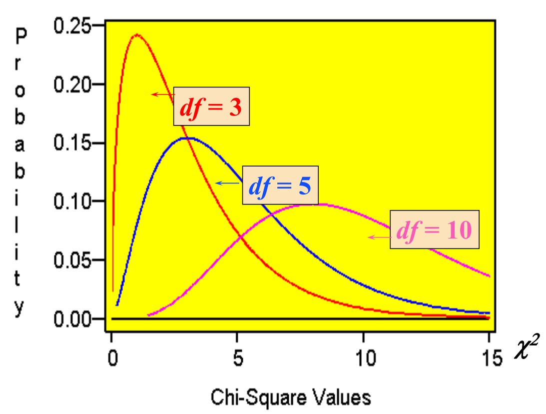
1

Characteristics of the Chi-Square Distribution

- The major characteristics of the chi-square distribution are:
  - It is positively skewed
  - It is non-negative
  - It is based on degrees of freedom
  - When the degrees of freedom change a new distribution is created

2

2



3

Goodness-of-Fit Test: Equal Expected Frequencies

- Let O_i and E_i be the observed and expected frequencies respectively for each category.
- H_0: there is no difference between Observed and Expected Frequencies
- H_a: there is a difference between Observed and Expected Frequencies
- The test statistic is: $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$
- The critical value is a chi-square value with (k-1) degrees of freedom, where k is the number of categories

4

4

EXAMPLE 1

- The following data on absenteeism was collected from a manufacturing plant. At the .01 level of significance, Can you support the claim that there is a difference in the absence rate by day of the week?

Day	Frequency
Monday	95
Tuesday	65
Wednesday	60
Thursday	80
Friday	100

5

5

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O_i	p_i
Mon	95	0.20
Tues	65	0.20
Wed	60	0.20
Thur	80	0.20
Fri	100	0.20
Total	**400**	**1**

6

6

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O_i	p_i	E_i
Mon	95	0.20	80
Tues	65	0.20	80
Wed	60	0.20	80
Thur	80	0.20	80
Fri	100	0.20	80
Total	**400**	**1**	**400**

7

7

EXAMPLE 1 *continued*

- Assume equal expected frequency: (95+65+60+80+100)/5=80

Day	O_i	p_i	E_i	(O-E)^2/E
Mon	95	0.20	80	2.8125
Tues	65	0.20	80	2.8125
Wed	60	0.20	80	5.0000
Thur	80	0.20	80	0.0000
Fri	100	0.20	80	5.0000
Total	**400**	**1**	**400**	**15.625**

8

8

EXAMPLE 1 *continued*

- H_o: There is no difference absenteeism due to day of the week.
- H_a: There is a difference absenteeism due to day of the week.
- H_o: $p_1=p_2=p_3=p_4=p_5$
- H_a: At least one proportion is different

- Test statistic: chi-square=Σ(O-E)^2/E=15.625
- Decision Rule: reject H_o if test statistic is greater than the critical value of 13.277. (4 df, α=.01)

- Conclusion: reject H_o and conclude that there is a difference absenteeism due to day of the week.

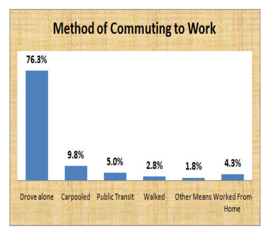
9

9

Goodness-of-Fit Test: Unequal Expected Frequencies

EXAMPLE 2

- In the 2010 United States census, data was collected on how people get to work -- their method of commuting.
- Suppose you wanted to know if people who live in the San Jose metropolitan area (Santa Clara County) commute with similar proportions as the United States.
- Design and conduct a hypothesis test at the 5% significance level.



10

10

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764			
Carpooled	105			
Public Transit	34			
Walked	20			
Other Means	30			
Worked from Home	47			
TOTAL	**1000**			

11

11

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764	0.763		
Carpooled	105	0.098		
Public Transit	34	0.050		
Walked	20	0.028		
Other Means	30	0.018		
Worked from Home	47	0.043		
TOTAL	**1000**	**1.000**		

12

12

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764	0.763	763	
Carpooled	105	0.098	98	
Public Transit	34	0.050	50	
Walked	20	0.028	28	
Other Means	30	0.018	18	
Worked from Home	47	0.043	43	
TOTAL	**1000**	**1.000**	**1000**	

13

13

EXAMPLE 2 *continued*

Method Of Commuting	Observed Frequency O_i	Expected Proportion p_i	Expected Frequency E_i	$\sum \frac{(O-E)^2}{E}$
Drive Alone	764	0.763	763	0.0013
Carpooled	105	0.098	98	0.5000
Public Transit	34	0.050	50	5.1200
Walked	20	0.028	28	2.2857
Other Means	30	0.018	18	8.0000
Worked from Home	47	0.043	43	0.3721
TOTAL	**1000**	**1.000**	**1000**	**16.2791**

14

14

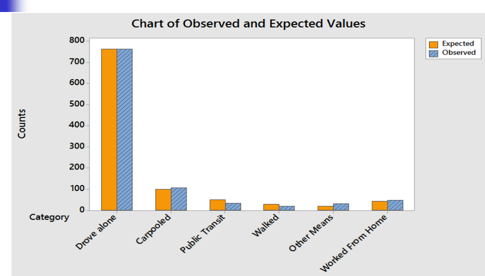
EXAMPLE 2 *continued*

- **Design:**
  Ho: p_1 = .763 p_2 = .098 p_3 = .050 p_4 = .028 p_5 = .018 p_6 = .043
  Ha: At least one p_i is different than what was stated in Ho
- α=.05
- Model: Chi-Square Goodness of Fit, df=5
- H_o is rejected if $\chi^2 > 11.071$
- **Data:**
- χ^2 = 16.2791, Reject Ho
- **Conclusion:**
  Workers in Santa Clara County do not have the same frequencies of method of commuting as workers in the entire United States.

15

15

EXAMPLE 2 *continued*



16

16

Explanatory/Response Models

- The remaining models covered in the course can be used for testing claims of the following form:
- Ho: There is no difference in the **Response Variable** due to the **Explanatory Variable**
- Ha: There is a difference in the **Response Variable** due to the **Explanatory Variable**
- If both the explanatory and categorical variables are categorical, then use the **Chi-square Test of Independence Model**

17

17

Chi-square Test of Independence

- Contingency table analysis is used to test whether two traits or variables are related.
- Each observation is classified according to two categorical variables (Explanatory and Response).
- Ha: The variables are dependent
- The *degrees of freedom* is equal to: (number of rows-1)(number of columns-1).
- The expected frequency is computed as: Expected Frequency = (row total)(column total)/grand total

18

18

EXAMPLE 3

- In May 2014, Colorado became the first state to legalize the recreational use of marijuana.
- A poll of 1000 adults were classified by gender and their opinion about legalizing marijuana
- At the .05 level of significance, can we conclude that gender and the opinion about legalizing marijuana for recreational use are dependent events?

Marijuana should be	Men	Women	Total
Legal	270	230	500
Not Legal	205	245	450
Unsure	25	25	50
Total	**500**	**500**	**1000**

19

19

Example 3 (continued)

- The **observed** is the reported data.

Observed Expected Chi-Sq	Men	Women	TOTAL
Legal	270	230	500
Not Legal	205	245	450
Unsure	25	25	50
TOTAL	500	500	1000

20

20

Example 3 (continued)

- The **observed** is the reported data.
- The **expected** is

$$\frac{Row\ Total\ \times Column\ Total}{Grand\ Total\ (n)}$$

Observed Expected Chi-Sq	Men	Women	TOTAL
Legal	270 250	230 250	500
Not Legal	205 225	245 225	450
Unsure	25 25	25 25	50
TOTAL	500	500	1000

21

21

Example 3 (continued)

- The **observed** is the reported data.
- The **expected** is

$$\frac{Row\ Total\ \times Column\ Total}{Grand\ Total\ (n)}$$

- **Chi-square** is

$$\frac{(Observed - Expected)^2}{Expected}$$

- Sum to get test statistic: χ^2 **= 6.756**

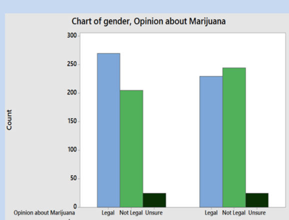
Observed Expected Chi-Sq	Men	Women	TOTAL
Legal	270 250 1.600	230 250 1.600	500
Not Legal	205 225 1.778	245 225 1.778	450
Unsure	25 25 0.000	25 25 0.000	50
TOTAL	500	500	1000

22

22

EXAMPLE 3 *continued*

```
Rows: Opinion about Marijuana
Columns: gender

1st value = Observed
2nd Value = Expected
3rd Value = Contribution to Chi-square

              men   women    All

Legal         270    230     500
              250    250
            1.600  1.600

Not Legal     205    245     450
              225    225
            1.778  1.778

Unsure         25     25      50
               25     25
            0.000  0.000

All           500    500    1000
```



23

23

EXAMPLE 3 *continued*

- Explanatory Variable: Gender Response Variable: Opinion
- H_o: There is no difference in Opinion due to gender.
  H_a: There is a difference in Opinion due to gender.
  H_o: Gender and Opinion are independent.
  H_a: Gender and Opinion are dependent.
- α=.05
- Model: Chi-Square Test for Independence, df=2
- H_o is rejected if χ^2 > 5.99
- **Data:** χ^2 = 6.756, Reject Ho
- **Conclusion:** Gender and opinion are dependent variables. Men are more likely to support legalizing marijuana for recreational use.

24

24

Inferential Statistics and Probability
a Holistic Approach

Chapter 12

One Factor Analysis of Variance (ANOVA)

1

1

ANOVA Definitions

- **Factor** – categorical variable that defines the populations.
- **Response** – variable that is being measured.
- **Levels** – the number of choices for the factor, represented by k
- **Replicates** – the sample size for each level, n_1, n_2, …, n_k.
- If $n_1 = n_2 = \ldots = n_k$, then the design is **balanced.**

- **Ho:** There is no difference in the mean <response in context> due to the <factor in context>.
- **Ha:** There is a difference in the mean <response in context> due to the <factor in context>.

2

2

Underlying Assumptions for ANOVA

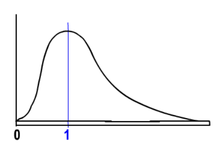
- The *F* distribution is also used for testing the equality of more than two means using a technique called analysis of variance (ANOVA). ANOVA requires the following conditions:
  - The populations being sampled are normally distributed.
  - The populations have equal standard deviations.
  - The samples are randomly selected and are independent.

3

3

Characteristics of F-Distribution

- There is a "family" of *F* Distributions.
- Each member of the family is determined by two parameters: the numerator degrees of freedom and the denominator degrees of freedom.
- *F* cannot be negative, and it is a continuous distribution.
- The *F* distribution is positively skewed.
- Its values range from 0 *to* ∞. As $F \rightarrow \infty$ the curve approaches the X-axis.



4

4

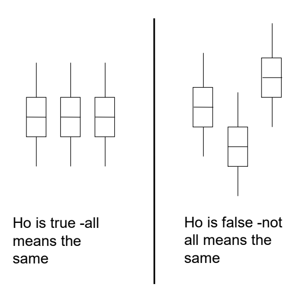
Analysis of Variance Procedure

- The Null Hypothesis: the population means are the same.
- The Alternative Hypothesis: at least one of the means is different.
- The Test Statistic: F=(between sample variance)/(within sample variance).
- Decision rule: For a given significance level α, reject the null hypothesis if *F* (computed) is greater than *F* (table) with numerator and denominator degrees of freedom.

5

5

ANOVA – Null Hypothesis



6

6

ANOVA NOTES

- If there are k populations being sampled (levels), then the df_{factor} = k-1
- If the sample size is n, then df_{error}= n-k
- The test statistic is computed by:F=[(SS_F)/(k-1)]/[(SS_E)/(n-k)].
- SS_F represents the factor (between) sum of squares.
- SS_E represents the error (within) sum of squares.
- Let T_C represent the column totals, n_c represent the number of observations in each column, and ΣX represent the sum of all the observations.
- These calculations are tedious, so technology is used to generate the **ANOVA table.**

7

7

Formulas for ANOVA

$$SS_{Total} = \Sigma\left(X^2\right) - \frac{(\Sigma X)^2}{n}$$

$$SS_{Factor} = \Sigma\left(\frac{T_c^2}{n_c}\right) - \frac{(\Sigma X)^2}{n}$$

$$SS_{Error} = SS_{Total} - SS_{Factor}$$

8

8

ANOVA Table

Source	SS	df	MS	F
Factor	SS_{Factor}	k-1	SS_F/df_F	MS_F/MS_E
Error	SS_{Error}	n-k	SS_E/df_E	
Total	SS_{Total}	n-1		

9

9

EXAMPLE

- Party Pizza specializes in meals for students. Hsieh Li, President, recently developed a new tofu pizza.
- Before making it a part of the regular menu she decides to test it in several of her restaurants. She would like to know if there is a difference in the mean number of tofu pizzas sold per day at the Cupertino, San Jose, and Santa Clara pizzerias for sample of five days.
- At the .05 significance level can Hsieh Li conclude that there is a difference in the mean number of tofu pizzas sold per day at the three pizzerias?

10

10

Example

	Cupertino	San Jose	Santa Clara	Total
	13	10	18	
	12	12	16	
	14	13	17	
	12	11	17	
			17	
T	51	46	85	182
n	4	4	5	13
Means	12.75	11.5	17	14
Σ^2	653	534	1447	2634

11

11

Example *continued*

$$SS_{Total} = 2634 - \frac{182^2}{13} = 86$$

$$SS_{Factor} = 2624.25 - \frac{182^2}{13} = 76.25$$

$$SS_{Error} = 86 - 76.25 = 9.75$$

12

12

Example 4 *continued*

ANOVA TABLE

Source	SS	df	MS	F
Factor	76.25	2	38.125	39.10
Error	9.75	10	0.975	
Total	86.00	12		

13

13

EXAMPLE 4 *continued*

- **Design:** H_o: $\mu 1=\mu 2=\mu 3$
  H_a: Not all the means are the same
- α=.05
- Model: One Factor ANOVA
- H_0 is rejected if F>4.10
- **Data:** Test statistic: F=[76.25/2]/[9.75/10]=39.1026
- H_0 is rejected.
- **Conclusion:** There is a difference in the mean number of pizzas sold at each pizzeria.

14

14

One-way ANOVA: Cupertino, San Jose, Santa Clara

```
Source  DF      SS      MS      F      P
Factor   2  76.250  38.125  39.10  0.000
Error   10   9.750   0.975
Total   12  86.000

S = 0.9874   R-Sq = 88.66%   R-Sq(adj) = 86.40%

                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level        N    Mean  StDev  --------+---------+---------+---------+-
Cupertino    4  12.750  0.957      (-----*----)
San Jose     4  11.500  1.291  (----*-----)
Santa Clara  5  17.000  0.707                          (----*----)
                               --------+---------+---------+---------+-
                                    12.0      14.0      16.0      18.0
```

15

15

Post Hoc Comparison Test

- Used for pairwise comparison
- Designed so the **overall** signficance level is 5%.
- Use technology.
- Refer to **Tukey Test** Material in the textbook.

16

16

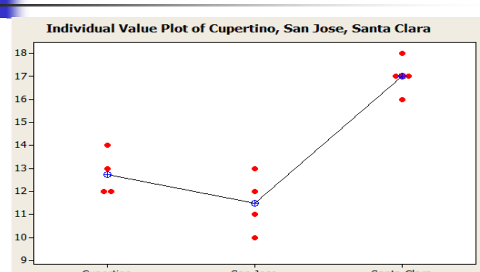
Post Hoc Comparison Test

```
Grouping Information Using Tukey Method

             N     Mean  Grouping
Santa Clara  5  17.0000  A
Cupertino    4  12.7500    B
San Jose     4  11.5000    B

Means that do not share a letter are significantly different.
```

17

17

Post Hoc Comparison Test



18

18

Example – Oranges & Orchards

- Valencia oranges were tested for juiciness at 4 different orchards. Eight oranges were sampled from each orchard, and the total ml of juice per 20 gms of orange was calculated.
- Test for a difference in juiciness due to orchards using alpha = .05
- Perform all the pairwise comparisons using Tukey's Test and an overall risk level of 5%.

Orchard A:	**Orchard B:**	**Orchard C:**	**Orchard D:**
11,13,12,14, 9,13,11,9	10,9,8,10, 11,12,7,8	13,15,14,11, 12,10,16,11	9,7,11,9, 9,11,10,8

19

19

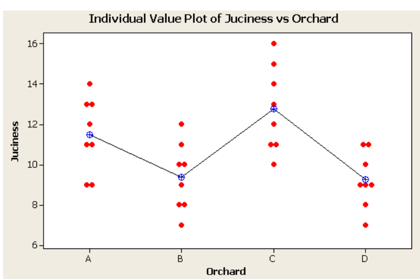
Example - Defintions

- Factor: Orchard (A, B, C or D)
- Response: Juiciness of orange
- Levels: k = 4
- Replicate: $n_A = n_B = n_C = n_D = 8$
- Design: Balanced
- Sample size: n = 8 + 8 + 8 + 8 = 32

20

20

Example – Value Plot



21

21

Example – Stats & ANOVA Table

Level	N	Mean	StDev
Orchard A	8	11.500	1.852
Orchard B	8	9.375	1.685
Orchard C	8	12.750	2.121
Orchard D	8	9.250	1.389

Source	DF	SS	MS	F	P
Factor	3	69.59	23.20	7.31	0.001
Error	28	88.88	3.17		
Total	31	158.47			

22

22

Example – Tukey Test Grouping

Grouping Information Using Tukey Method

	N	Mean	Grouping
Orchard C	8	12.750	A
Orchard A	8	11.500	A B
Orchard B	8	9.375	B
Orchard D	8	9.250	B

Means that do not share a letter are significantly different.

23

23

Inferential Statistics and Probability
a Holistic Approach

Chapter 13
Correlation and Linear Regression

1

1

Mathematical Model

- You have a small business producing custom t-shirts.
- Without marketing, your business has revenue (sales) of $1000 per week.
- Every dollar you spend marketing will increase revenue by 2 dollars.
- Let variable X represent amount spent on marketing and let variable Y represent revenue per week.
- Write a mathematical model that relates X to Y

2

2

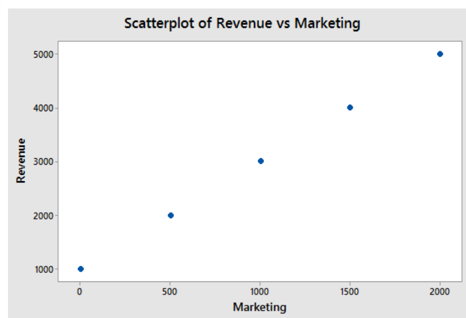
Mathematical Model - Table

X=marketing	Y=revenue
$0	$1000
$500	$2000
$1000	$3000
$1500	$4000
$2000	$5000

3

3

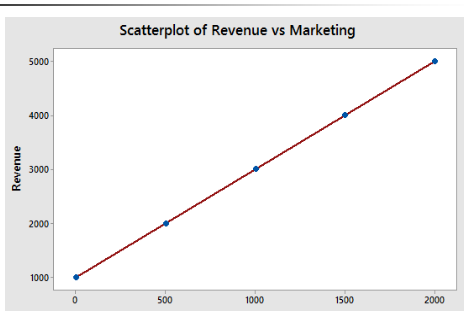
Mathematical Model - Scatterplot



4

4

Mathematical Model - Linear



5

5

Mathematical Linear Model

Linear Model	**Example**
$Y = \beta_0 + \beta_1 X$	$Y = 1000 + 2X$
Y: Dependent Variable	Y: Revenue
X: Independent Variable	X: Marketing
β_0: Y-intercept	β_0: $1000
β_1: Slope	β_1: $2 per $1 marketing

6

6

Statistical Model

- You have a small business producing custom t-shirts.
- Without marketing, your business has **expected** revenue (sales) of $1000 per week.
- Every dollar you spend marketing will increase revenue by an **expected value** of 2 dollars.
- Let variable X represent amount spent on marketing and let variable Y represent revenue per week.
- Let ε represent the difference between Expected Revenue and Actual Revenue (Residual Error)
- Write a statistical model that relates X to Y

7

7

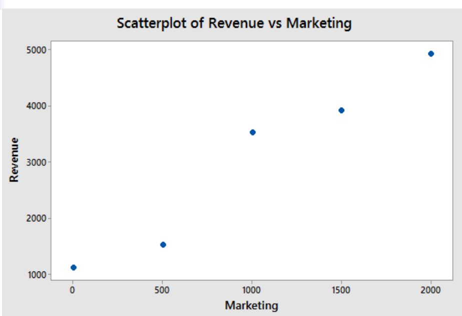
Statistical Model - Table

X=Marketing	Expected Revenue	Y=Actual Revenue	ε=Residual Error
$0	$1000	$1100	+$100
$500	$2000	$1500	-$500
$1000	$3000	$3500	+$500
$1500	$4000	$3900	-$100
$2000	$5000	$4900	-$100

8

8

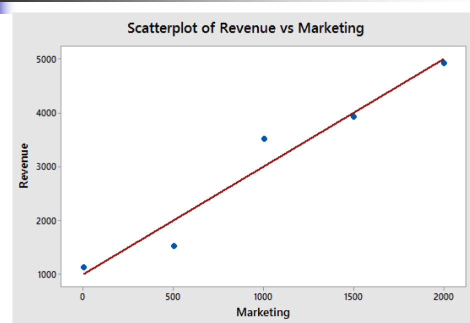
Statistical Model - Scatterplot



9

9

Statistical Model - Linear



10

10

Statistical Linear Model

Regression Model

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

Y: *Dependent Variable*

X: *Independent Variable*

β_0: *Y-intercept*

β_1: *Slope*

ε: *Normal*$(0, \sigma)$

Example

$Y = 1000 + 2X + \varepsilon$

Y: Revenue

X: *Marketing*

β_0: $1000

β_1: $2 *per* $1 *marketing*

11

11

Regression Analysis

- Purpose: to determine the regression equation; it is used to predict the value of the dependent response variable (Y) based on the independent explanatory variable (X).
- Procedure:
  - select a sample from the population
  - list the paired data for each observation
  - draw a scatter diagram to give a visual portrayal of the relationship
  - determine the regression equation.

12

12

Simple Linear Regression Model

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

$Y:$ *Dependent Variable*

$X:$ *Independent Variable*

$\beta_0:$ *Y − intercept*

$\beta_1:$ *Slope*

$\varepsilon:$ *Normal* $(0, \sigma)$

13

13

Estimation of Population Parameters

- From sample data, find statistics that will estimate the 3 population parameters
- Slope parameter
  - b_1 will be an estimator for β_1
- Y-intercept parameter
  - b_o will be an estimator for β_o
- Standard deviation
  - s_e will be an estimator for σ

14

14

Regression Analysis

- the regression equation: $\hat{Y} = b_0 + b_1 X$, where:
- $\hat{Y}$ is the average predicted value of *Y* for any *X*.
- b_0 is the Y-intercept, or the estimated *Y* value when *X=0*
- b_1 is the slope of the line, or the average change in $\hat{Y}$ for each change of one unit in *X*
- the least squares principle is used to obtain b_1 *and* b_0

$$SSX = \Sigma X^2 - \tfrac{1}{n}(\Sigma X)^2 \qquad b_1 = \frac{SSXY}{SSX}$$

$$SSY = \Sigma Y^2 - \tfrac{1}{n}(\Sigma Y)^2$$

$$SSXY = \Sigma XY - \tfrac{1}{n}(\Sigma X \cdot \Sigma y) \qquad b_0 = \overline{Y} - b_1 \overline{X}$$

15

15

Assumptions Underlying Linear Regression

- For each value of X, there is a group of Y values, and these Y values are *normally distributed.*
- The *means* of these normal distributions of Y values all lie on the straight line of regression.
- The *standard deviations* of these normal distributions are equal.
- The Y values are statistically independent. This means that in the selection of a sample, the Y values chosen for a particular X value do not depend on the Y values for any other X values.

16

16

Example

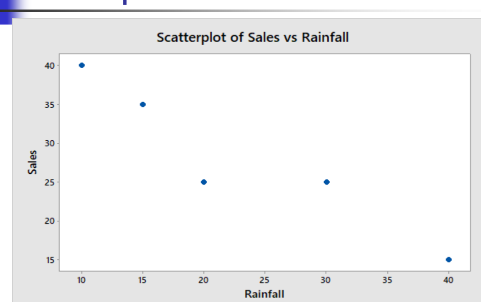
- X = Average Annual Rainfall (Inches)
- Y = Average Sale of Sunglasses/1000
  - Make a Scatterplot
  - Find the least square line

X	10	15	20	30	40
Y	40	35	25	25	15

17

17

Example *continued*



18

18

Example *continued*

	X	Y	X^2	Y^2	XY
	10	40	100	1600	400
	15	35	225	1225	525
	20	25	400	625	500
	30	25	900	625	750
	40	15	1600	225	600
Σ	115	140	3225	4300	2775

19

19

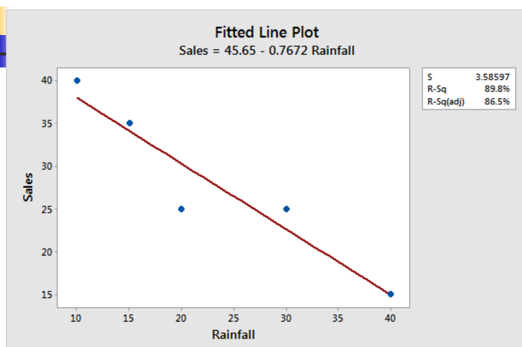
Example *continued*

- Find the Regression line
  - SSX = 580
  - SSY = 380
  - SSXY = -445
  - b_1 = -.767
  - b_0 = 45.647
  - $\hat{Y}$ = 45.647 - .767X

20

20

Minitab: Stat>Regression>Fitted Line Plot



21

21

Determining Regression Line

- Residual error for any observation is the difference between the observed and expected values of Y|X.
- For a given point (X,Y), $\hat{Y} = b_0 + b_1 X$
- Residual error for this point = $Y - \hat{Y}$
- We then minimize total error by combing all residuals
- Regression Line minimizes SSE = the sum of the squared residual errors

$$SSE = \sum \left(Y - \hat{Y}\right)^2$$

22

22

Example *continued*

- Find SSE and the
  - SSR = 341.422
  - SSE = 38.578

x	y	ŷ	y - ŷ	$(y - \hat{y})^2$
10	40	37.97	2.03	4.104
15	35	34.14	0.86	0.743
20	25	30.30	-5.30	28.108
30	25	22.63	2.37	5.620
40	15	14.96	0.04	0.002
			Total	38.578

23

23

Interpreting Regression Line

- Slope is the change in Y per the change in X.
- Example
- $\hat{Y}$ = 45.647 - .767X

Each increase of 1 inch of rainfall decreases Sales by 0.767

24

24

Hypothesis Testing in Simple Linear Regression

- The following Tests are equivalent:
  - H_0: There is no difference in Response(Y) due to Explanatory(X)
  - H_a: There is a difference in Response(Y) due to Explanatory(X)
  - H_0: X and Y are uncorrelated
  - H_a: X and Y are correlated
  - H_0: $\beta_1 = 0$
  - H_a: $\beta_1 \neq 0$

25

25

Hypothesis Testing Example

- H_0: There is no difference in Sales of Sunglasses due to Rainfall
- H_a: There is a difference in Sales of Sunglasses due to Rainfall
- H_0: Sales of Sunglasses and Rainfall are uncorrelated
- H_a: Sales of Sunglasses and Rainfall are correlated
- H_0: $\beta_1 = 0$
- H_a: $\beta_1 \neq 0$

26

26

ANOVA Table for Simple Linear Regression

Source	SS	df	MS	F
Regression	SSR	1	SSR/dfR	MSR/MSE
Error/Residual	SSE	n-2	SSE/dfE	
TOTAL	SSY	n-1		

27

27

Example *continued*

- Test the Hypothesis H_o: $\beta_1 = 0$, α=5%

Source	SS	df	MS	F	p-value
Regression	341.422	1	341.422	26.551	0.0142
Error	38.578	3	12.859		
TOTAL	380.000	4			

- Reject Ho p-value < α

28

28

The Standard Error of Estimate

- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression
- The formulas that are used to compute the standard error:

$$SSR = b_1 \cdot SSXY$$

$$SSE = \sum (Y - \hat{Y})^2 = SSY - SSR$$

$$MSE = SSE/(n-2)$$

$$s_e = \sqrt{MSE}$$

29

29

Example *continued*

- Find SSE and the standard error:
  - SSR = 341.422
  - SSE = 38.578
  - MSE = 12.859
  - s_e = 3.586

x	y	ŷ	y - ŷ	$(y - \hat{y})^2$
10	40	37.97	2.03	4.104
15	35	34.14	0.86	0.743
20	25	30.30	-5.30	28.108
30	25	22.63	2.37	5.620
40	15	14.96	0.04	0.002
			Total	38.578

30

30

Correlation Analysis

- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated. "Effect"
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable. "Cause?" (Maybe!)

31

31

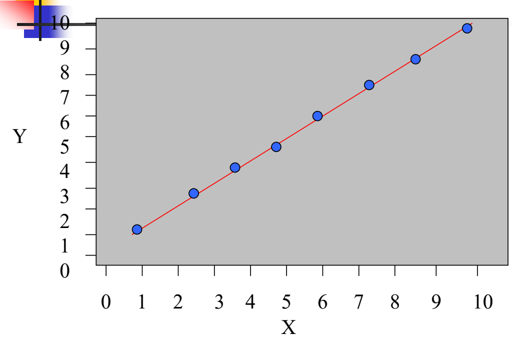
The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the **strength** of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1.00 to 1.00.
  - Values of -1.00 or 1.00 indicate perfect and strong correlation.
  - Values close to 0.0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

32

32

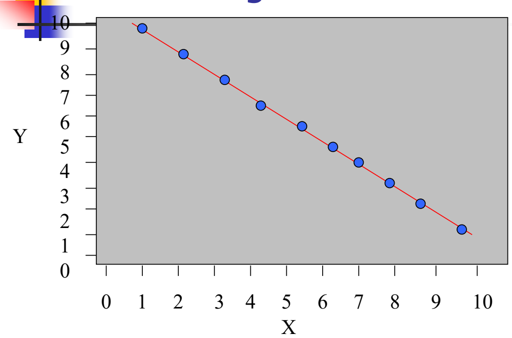
Perfect Positive Correlation



33

33

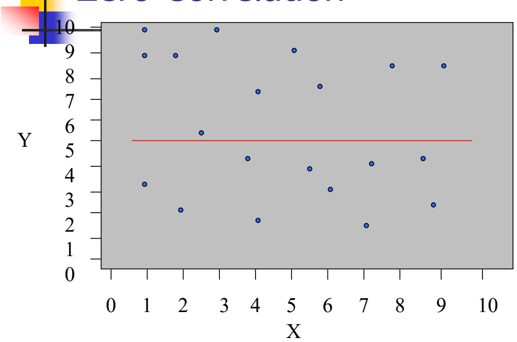
Perfect Negative Correlation



34

34

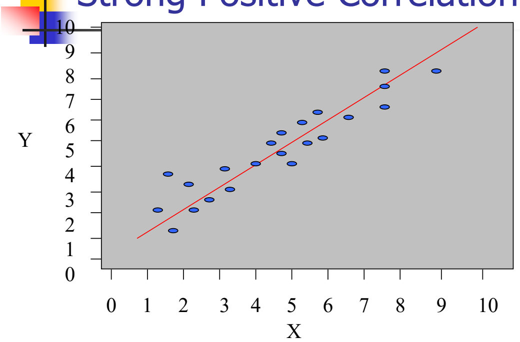
Zero Correlation



35

35

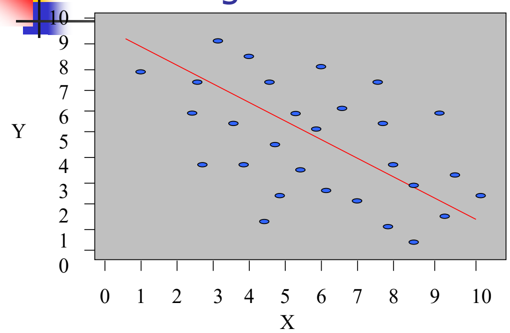
Strong Positive Correlation



36

36

Weak Negative Correlation



37

37

r^2: Coefficient of Determination

- r^2 is the proportion of the total variation in the dependent variable Y that is explained or accounted for by the variation in the independent variable X.

$$r^2 = \frac{SSR}{SSTotal}$$

38

38

Example *continued*

$$r^2 = \frac{341.422}{380.000} = 0.8985 = 89.85\%$$

- 89.85% of the variability of Sales of Sunglasses is explained by Rainfall
- 10.15% of the variability of Sales of Sunglasses is unexplained

39

39

Example *continued*

X	Y	X^2	Y^2	XY
10	40	100	1600	400
15	35	225	1225	525
20	25	400	625	500
30	25	900	625	750
40	15	1600	225	600
115	140	3225	4300	2775

- SSX = $3225 - 115^2/5$ = 580
- SSY = $4300 - 140^2/5$ = 380
- SSXY= 2775 - (115)(140)/5 = -445

40

40

Confidence Interval

- The confidence interval for the mean value of Y for a given value of X is given by:

$$\hat{Y} \pm t \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{SSX}}$$

- Degrees of freedom for t =n-2

41

41

Prediction Interval

- The prediction interval for an individual value of Y for a given value of X is given by:

$$\hat{Y} \pm t \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{SSX}}$$

- Degrees of freedom for t =n-2

42

42

Example *continued*

- Find a 95% Confidence Interval for Sales of Sunglasses when rainfall = 25 inches.
- Find a 95% Prediction Interval for Sales of Sunglasses when rainfall = 25 inches.

43

43

Example – Minitab output

- Sales = 45.65 - 0.767 Rainfall
- Variable Setting
- Rainfall 25
- Fit SE Fit 95% CI 95% PI
- 26.4655 1.63111 (21.2746, 31.6564) (13.9282, 39.0028)

44

44

Example *continued*

- 95% Confidence Interval

$$22.63 \pm 6.60$$

- 95% Confidence Interval

$$22.63 \pm 13.18$$

45

45

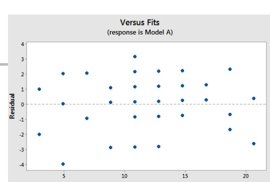
Residual Analysis

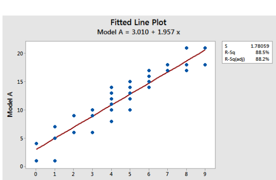
- Residuals for Simple Linear Regression
  - The residuals should represent a linear model.
  - The standard error (standard deviation of the residuals) should not change when the value of X changes.
  - The residuals should follow a normal distribution.
  - Look for any potential extreme values of X.
  - Look for any extreme residual errors

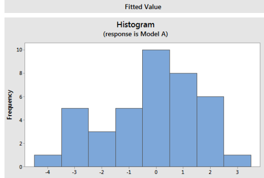
46

46

Example A



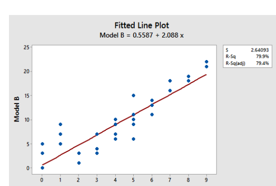


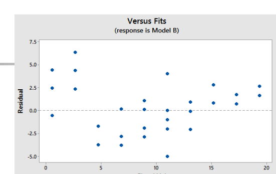


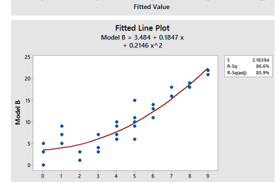
47

47

Example B



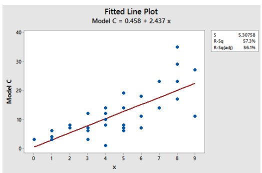


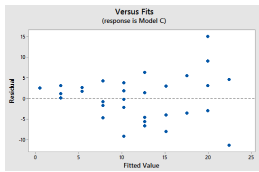


48

48

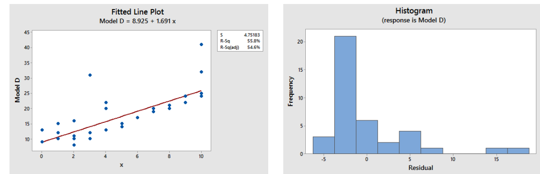
Example C





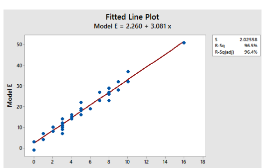
49

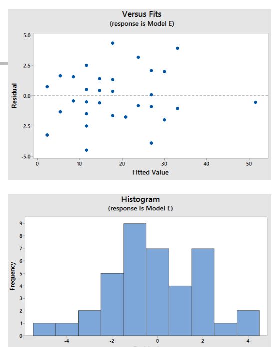
Example D



50

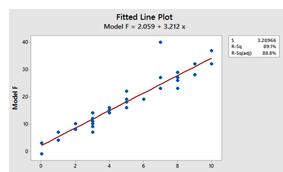
Example E

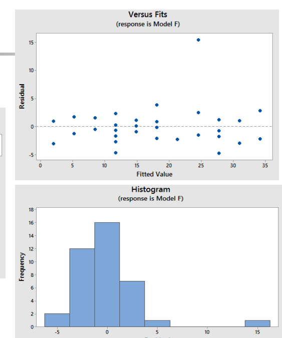




51

Example F





52

Using Minitab to Run Regression

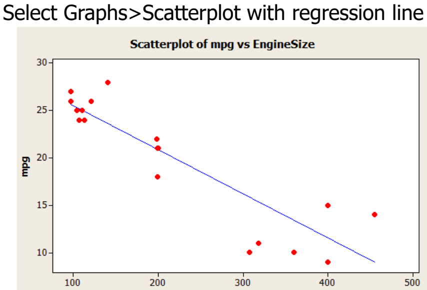
- Data shown is engine size in cubic inches (X) and MPG (Y) for 20 cars.

x	y		x	y
400	15		104	25
455	14		121	26
113	24		199	21
198	22		360	10
199	18		307	10
200	21		318	11
97	27		400	9
97	26		97	27
110	25		140	28
107	24		400	15

53

Using Minitab to Run Regression

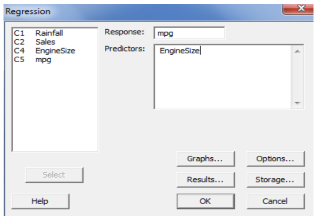
Select Graphs>Scatterplot with regression line



54

Using Minitab to Run Regression

Select Statistics>Regression>Regression, then choose the Response (Y-variable) and model (X-variable)

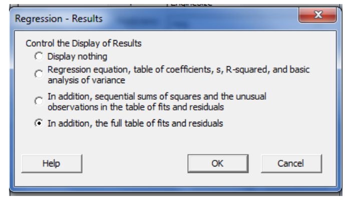


55

55

Using Minitab to Run Regression

Click the results box, and choose the fits and residuals to get all predictions.



56

56

Using Minitab to Run Regression

The results at the beginning are the regression equation, the intercept and slope, the standard error of the residuals, and the r^2

```
The regression equation is
mpg = 30.2 - 0.0466 EngineSize


Predictor        Coef   SE Coef      T      P
Constant       30.203     1.361  22.20  0.000
EngineSize  -0.046598  0.005378  -8.66  0.000


S = 2.95688   R-Sq = 80.7%   R-Sq(adj) = 79.6%
```

57

57

Using Minitab to Run Regression

Next is the ANOVA table, which tests the significance of the regression model.

```
Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  656.42  656.42  75.08  0.000
Residual Error  18  157.38    8.74
Total           19  813.80
```

58

58

Using Minitab to Run Regression

Finally, the residuals show the potential outliers.

Obs	EngineSize	mpg	Fit	SE Fit	Residual	St Resid
1	400	15.000	11.564	1.167	3.436	1.26
2	455	14.000	9.001	1.421	4.999	1.93
3	113	24.000	24.937	0.880	-0.937	-0.33
4	198	22.000	20.976	0.673	1.024	0.36
5	199	18.000	20.930	0.672	-2.930	-1.02
6	200	21.000	20.883	0.671	0.117	0.04
7	97	27.000	25.683	0.939	1.317	0.47
8	97	26.000	25.683	0.939	0.317	0.11
9	110	25.000	25.077	0.891	-0.077	-0.03
10	107	24.000	25.217	0.902	-1.217	-0.43
11	104	25.000	25.357	0.913	-0.357	-0.13
12	121	26.000	24.565	0.853	1.435	0.51
13	199	21.000	20.930	0.672	0.070	0.02
14	360	10.000	13.427	0.998	-3.427	-1.23
15	307	10.000	15.897	0.807	-5.897	-2.07R
16	318	11.000	15.385	0.842	-4.385	-1.55
17	400	9.000	11.564	1.167	-2.564	-0.94
18	97	27.000	25.683	0.939	1.317	0.47
19	140	28.000	23.679	0.792	4.321	1.52
20	400	15.000	11.564	1.167	3.436	1.26

59

59

Using Minitab to Run Regression

- Find a 95% confidence interval for the **expected** MPG of a car with an engine size of 250 ci.
- Find a 95% prediction interval for the **actual** MPG of a car with an engine size of 250 ci.

```
mpg = 30.20 - 0.04660 EngineSize


Variable    Setting
EngineSize      250


    Fit    SE Fit          95% CI               95% PI
18.5533  0.679201  (17.1264, 19.9803)  (12.1793, 24.9273)
```

60

60