

Math 127

Intro To Statistics

Cecil College

Course Notes

Joseph J. Kupresanin

Author's Note:

This work was written for use by Cecil College students when I was granted a sabbatical during the Fall 2013 semester. Though we've kept the material essentially in-house, those outside the Cecil community are welcome to use these materials as they see fit.

I am very proud of the final product, and if you use these materials, I hope you learn enough statistics to help you excel in other college courses and throughout your career.

With the gracious help of Kim Sheppard, we recorded 36 hours of videos over the course of 9 months to accompany the White Notes for Math 127. Those videos will live on YouTube for the indefinite future.

It is my intention that, as long as Cecil College chooses to use these materials for Math 127, this content remains free of charge to the Cecil College students who register for Math 127, and that it will be distributed by the first day of class by the instructor in print form.

To be sure, I am very proud to have worked at a college that supported my sabbatical and provides these course materials to students – in fact, this book is my proudest professional accomplishment, and I truly hope you enjoy the course.

The Math 127 course materials include the following:

- 36 hours of statistics videos found on YouTube
- The White (at-home) Notes, which accompany the video lectures
- The Yellow (in-class) Notes, which accompany the instructor lectures
- The Pink formula sheets
- A plethora of digital course material found on the Blackboard page
- Over 700 StatCrunch online homework problems found on the Blackboard page

All this content is available to use under a Creative Commons license. Please consider the following:

You are free to copy and redistribute this material, just please give appropriate credit. You may not use the material for commercial purposes. This material may not be sold. No derivatives. If you transform or build upon the material, you may not distribute the modified material.



I wish you much success in the course!

Professor Kupe
October 23, 2017

Table of Contents

Lesson	Unit	Topic	Page Number
0	I	Two Introductory Examples	1
1	I	Working With Data	7
2	I	Graphing Data	12
3	I	Summarizing Categorical Variables	24
4	I	Summarizing Quantitative Variables	34
5	I	Collecting Data With Observational Studies or Surveys	55
6	I	Collecting Data With Designed Experiments	68
7	II	The Difference Between Data and Models	77
8	II	Linear Models – Graphs and Summary Statistics	85
9	II	Linear Models – Least Squares Regression	100
10	II	Nonlinear Models	128
11	II	Probability	134
12	II	The Normal Model	155
13	II	The Binomial Model	179
14	III	Modeling the Sample Proportion	196
15	III	Confidence Intervals for Proportions	211
16	III	Hypothesis Tests for Proportions	226
17	III	Two-Sample Proportion Methods	253
18	III	Modeling the Sample Mean	264
19	III	Confidence Intervals for Means	274
20	III	Hypothesis Tests for Means	288
21	III	Two-Sample Mean Methods	298

Lesson 0

Page 1: Cecil College Fry Project ♦ 6 200 Countries in 200 Years

Lesson 1

Page 7: Who, What ♦ 9 When, Where, Why, How ♦ 10 Quantitative vs. Categorical ♦ 11 Identifiers

Lesson 2

Page 12: Lady with a Data Visualization ♦ 13 Pie Charts, Bar Plots ♦ 15 Stacked Bar Plots ♦ 16 Split Bar Plots ♦ 19 Histograms ♦ 20 Shape, Symmetry, Skewness, Modes ♦ 22 Center, Spread, Unusual Features

Lesson 3

Page 24: Categorical Variables: Proportions & Percentages ♦ 26 Contingency Tables ♦ 31 Independence Versus Dependence for Categorical Variables

Lesson 4

Page 34: Mean and Median ♦ 38 Quartiles ♦ 39 Percentiles ♦ 40 Range and IQR ♦ 41 Standard Deviation ♦ 44 Inventing Datasets with Certain Characteristics ♦ 45 Five-Number Summary ♦ 46 Outlier Rule-of-Thumb ♦ 47 Boxplots ♦ 48 Dealing with Outliers ♦ 49 Shifting and Scaling ♦ 52 Z-Scores ♦ 53 Unusual Observations

Lesson 5

Page 55: Battling Bad Science ♦ 57 Population vs. Sample ♦ 58 Observational Studies ♦ 60 Three Main Ideas of Sampling ♦ 62 Sampling Definitions ♦ 66 Post Office Example ♦ 67 Warehouse Example

Lesson 6

Page 68: Designed Experiments, Main Points ♦ 70 MLB Steroid Example ♦ 71 Designed Experiments Definitions ♦ 74 Poker Example ♦ 75 Butterfly Example

Lesson 7

Page 77: The Difference Between Data and Models IQ Example ♦ 82 Haagen-Dazs Example

Lesson 8

Page 85: Direct Loans Example ♦ 86 Roles of Variables, Explanatory vs. Response ♦ 87 Scatterplots ♦ 88 Form, Direction, Strength, Unusual Features ♦ 90 Correlation Coefficient ♦ 95 Facts About Correlation ♦ 97 Hans Rosling Correlations ♦ 99 Lurking Variables

Lesson 9

Page 100: Determining the Best Line Using Residuals ♦ 101 Principle of Least Squares ♦ 102 The Equation for the Linear Model ♦ 103 Interpreting Slope ♦ 104 Interpreting the y -Intercept ♦ 105 Slope and Intercept Examples ♦ 106 Formulas for the Slope and Intercept ♦ 108 Neighborhood Example ♦ 110 Making Predictions ♦ 111 Extrapolation ♦ 112 Temperature Example ♦ 114 Checking Conditions ♦ 115 Residual Plots ♦ 118 Standard Error of the Residuals ♦ 121 R-Sq ♦ 122 More on R-Sq ♦ 125 Library Book Example

Lesson 10

Page 128: Nonlinear Relationships ♦ 130 Exponential Functions ♦ 132 Quadratic Functions

Lesson 11

Page 134: Why Study Probability? ♦ 135 Probability Definitions ♦ 137 Two Probability Examples ♦ 140 Probability Rules Part 1 ♦ 147 Probability Rules Part 2 ♦ 149 Conditional Probability ♦ 151 Independence ♦ 153 Benefits Example

Lesson 12

Page 155: The Normal Model ♦ 158 The Empirical Rule ♦ 160 Linking Z-Scores to the Normal Model ♦ 161 The Standard Normal Model ♦ 164 Using StatCrunch to Solve Normal Model Problems ♦ 176 Checking Data for Normality

Lesson 13

Page 178: The Binomial Model ♦ 182 Binomial Probabilities ♦ 183 The Mean and Standard Deviation for a Binomial Model ♦ 188 Using StatCrunch to Solve Binomial Model Problems ♦ 191 The Normal Approximation to the Binomial Model

Lesson 14

Page 196: Modeling the Sample Proportion, Smoking Example ♦ 198 Smoking Example Simulation ♦ 200 The Mean and Standard Deviation of the Sample Proportion ♦ 201 Conditions and Assumptions ♦ 205 Problem Gamblers Example ♦ 207 Gun Ownership Example

Lesson 15

Page 211: Confidence Interval Main Ideas ♦ 214 Confidence Interval for a Proportion ♦ 215 Finding Critical Z Values ♦ 217 Problem Gamblers Example ♦ 218 Cecil College Gym Example ♦ 219 Facts About Confidence Intervals ♦ 223 Determining Sample Size

Lesson 16

Page 226: Hypothesis Testing Main Ideas ♦ 230 Official Steps to Run a One-Proportion Z-Test ♦ 232 P-Values ♦ 233 College Cheating Example ♦ 235 High School Attendance Example ♦ 237 Making Errors When Hypothesis Testing ♦ 243 Two-Tailed Tests ♦ 245 Fixed-Alpha Level Tests ♦ 247 Using Confidence Intervals to Run Tests ♦ 250 Terminology Review

Lesson 17

Page 253: Tests and Intervals for the Difference in Two Proportions ♦ 259 Official Steps and Formulas ♦ 260 Cecil vs. Harford Example ♦ 262 Retirement Example

Lesson 18

Page 264: Modeling the Sample Mean, IQ Example ♦ 267 Requirements to Model the Sample Mean with a Normal Model ♦ 268 Official Steps and Formulas ♦ 269 Pregnancy Example ♦ 272 Home Value Example ♦ 273 Grading Tests Example

Lesson 19

Page 274: Confidence Intervals for Means ♦ 275 The Big Problem and the Big Solution ♦ 276 Gossett ♦ 277 The Student's t Distribution ♦ 279 Official Steps for a Confidence Interval for a Mean ♦ 280 Finding Critical t Values on StatCrunch ♦ 282 Cecil College IQ Example ♦ 285 Kupe's Bowling Scores Example ♦ 286 Determining Sample Size

Lesson 20

Page 288: Official Steps to Run a t -Test ♦ 289 Credit Hours Example ♦ 292 Continental Tires Example ♦ 295 Radon Detectors Example ♦ 296 Sprinkler Example

Lesson 21

Page 298: Independent vs. Dependent Samples ♦ 299 Official Steps: Two-Sample t Test ♦ 301 CCLA Example ♦ 305 Textbook Example ♦ 307 Testing Dependent Samples

Lesson 0: Two Introductory Examples

You are taking a statistics class, not a math class! This class is about using data to make important decisions. The ideas learned in Statistics show up in many other college courses and you will encounter data in your careers and personal lives. We introduce the course with two examples.

Example 1: Cecil College French Fry Project

- On McDonald's website, we learn that each order of small fries is supposed to weigh 71 grams.
- The main point of the project was to check up on McDonald's claim. Are the local restaurants doing what they're supposed to do?
- **Question:** Can we expect each order of fries to weigh exactly 71 grams? _____
- Fry weights will _____, order to order. In one way, the whole field of Statistics exists because of this _____.

A few semesters back, the students in Math 127 visited three local McDonald's restaurants and purchased orders of small French fries.

In each section of Math 127, students decided amongst themselves which location to visit.

Then, for example, the “**Rising Sun**” group in one class was instructed to divvy up the days and times they went to buy their fries. This happened for all sections of Math 127 and all three locations. Why would this be important?

The fries were brought directly to Cecil College and weighed on the same digital scale. Why would this be important?

- Now, when conducting a study, the researchers are interested in the entire **population**. Since this project was about our three local McDonald's and their orders of small fries, what is the population for this problem? (Put yourself in the position of "Quality Control Manager" for these three restaurants – what is the population of interest?)

Population: _____

- In nearly all situations, it's really hard to analyze population data. We just won't take the time or spend the money to weigh every order of fries that we sell at McDonald's. The solution is to take a **sample** and analyze all of the characteristics contained in the sample. For this example, identify the sample.

Sample: _____

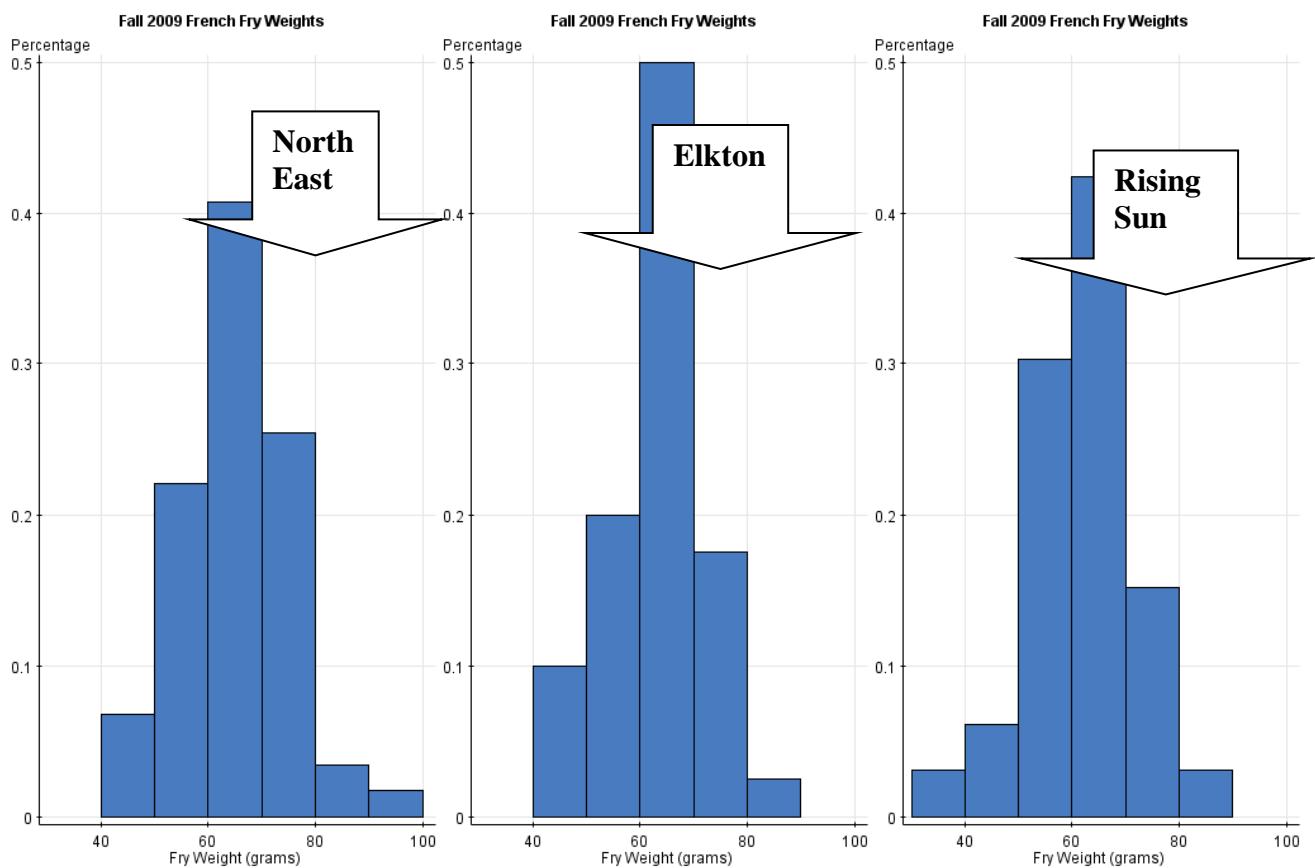
Draw a Diagram of Population vs. Sample for McDonald's:

- Students visited their location, returned to Cecil, weighed the fries, and recorded some information in a notebook left in the Math Lab. That information is called _____.

- Raw data is sort of useless. It was entered into StatCrunch. Here is part of it:

Row	Date	Time	Location	Weight	var5	va
1	Jan 28	11:20 a.m.	Elkton	54.7		
2	Jan 28	11:25 a.m.	elkton	70.5		
3	1/28	12	Rising Sun	54.1		
4	1/28	12:15 pm	Elkton	66.5		
5	1/28	12:18 p.m.	NE	83		
6	1/28	5:20 p.m.	Elkton	74.7		
7	1/28	1144 am	North East	70.6		
8						

- In Statistics, we take raw data and summarize it by making pictures and calculating summary statistics. Here are some graphs.



Comments:

- The first step is always make a picture. Then, once we have a sense of the distribution of our data values, we can move on to calculating the best summary statistics.
- Many important summary statistics have been calculated below. Define and discuss.

Column	<i>n</i>	Mean	Median	Min	Max	<i>Q₁</i>	<i>Q₃</i>	Std. Dev.
North East	59	65.42	65.11	43.44	98.36	59.21	71.81	10.20
Elkton	40	63.00	64.17	40.26	89	59.15	67.85	9.70
Rising Sun	33	62.64	63.31	39.34	87.19	56.25	69.28	9.84

A Few Final Questions About McDonald's

1. Remember that small fries are supposed to weigh 71 grams. Is McDonald's meeting its claim? Why or why not?
 2. What would happen to the shapes of the graphs and the values of the statistics if we took a new sample?
 3. To answer our research questions, though, in practice, how many samples will we take?
 4. **Tricky but important:** If the true mean weight of a small order of fries is really 71 grams, what percentage of orders will weigh in below 71 grams?¹

¹ This tells us that just because the sample mean for a particular location is under 71 grams, it does not necessarily mean that the location is under-filling the fries. The sample mean would have to be convincingly below 71 grams before we concluded McDonald's is under-filling its orders.

Example 2: 200 Countries in 200 Years

Here is an interesting 4-minute video on life expectancies, wealth, and the changes that took place for 200 countries during the past 200 years. The main point is to appreciate the power of good graphs and to get excited about the power of data.

Blackboard Link:

1. What graph was made by Hans Rosling? _____

2. Which variable was on the y -axis? _____.

In statistics, we call this the _____.

3. Which variable was on the x -axis? _____.

In statistics, we call this the _____.

4. Based on the analysis, what conclusions can be made about the relationship between the variables?

Lesson 1: Working With Data

Blackboard Video 1-A

“Statistics Before Calculus”

Blackboard Video 1-B

- We always begin a statistical investigation by understanding the context or story behind the data.

We need to think like a journalist – noting all the important facts about the data we have collected or have found. For every problem, be sure you can address the *Six W's*:

- The **Who** – Hardest to put into words, the **Who** is what you're collecting data on.

For the McDonald's example, identify the **Who**: _____

For the 200 Countries / 200 Years video, identify the **Who**: _____

When thinking about a dataset, the **Who** correspond to the _____ or the _____ or the _____.

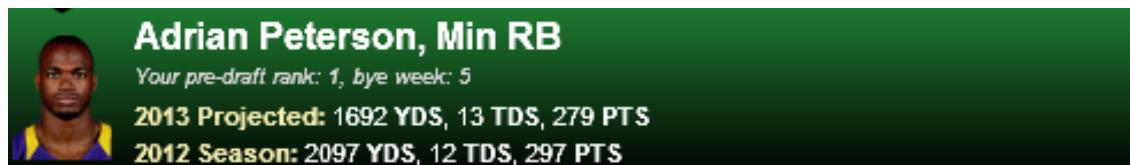
- The **What** – These are the variables. In Statistics, the variables are the characteristics recorded for each individual in our dataset.

For the McDonald's example, identify the **What**:

For the 200 Countries / 200 Years video, identify the **What**:

- When we analyze data, we typically will use the software called StatCrunch. It is important to understand exactly how data is entered into software.
- Each _____ corresponds to an individual. Put the **Who** in _____.
- Each _____ corresponds to a variable. Put the **What** in _____.

Example: On ESPN.com, fantasy football data is presented in the following table. Identify the **Who** and the **What**.



The screenshot shows a player profile for Adrian Peterson, a Minnesota RB. It includes his photo, pre-draft rank (1), bye week (5), and projected stats for 2013 (1692 YDS, 13 TDS, 279 PTS) and 2012 Season (2097 YDS, 12 TDS, 297 PTS). Below the profile is a navigation bar with tabs for Players, Draft Summary, RECOMMENDS, and Rules. Underneath are filters for Position (ALL) and Team (ALL). A table lists the top 6 running backs by rank, showing their name, team, position, and various performance metrics like BYE, PTS, COM, ATT, PAYD, PATD, and INT.

RANK	PLAYER	BYE	PTS	COM	ATT	PAYD	PATD	INT
1.	Adrian Peterson Min, RB	5	279	0	0	0	0	0
2.	Arian Foster Hou, RB	8	261	0	0	0	0	0
3.	Marshawn Lynch Sea, RB	12	233	0	0	0	0	0
4.	Ray Rice Bal, RB	8	232	0	0	0	0	0
5.	Doug Martin TB, RB	5	226	0	0	0	0	0
6.	Jamaal Charles KC, RB	10	228	0	0	0	0	0

The **Who**:

The **What**:

Blackboard Video 1-C

- The **Who** and the **What** are the most important features to understand, but we must also be aware _____ and _____ the data was collected, _____ it was collected, and _____ it was collected.
- If data was collected years ago, it might now be _____.
- If data was collected unscrupulously, our conclusions will be _____.

Example: Here are a few of Professor Kupe's financial transactions, recorded on Mint.com.

<input type="checkbox"/>	Date	Description	Category		Amount
▼ Pending (4)					
<input type="checkbox"/>	SEP 1	Megan's Books	Uncategorized		-\$49.96
<input type="checkbox"/>	AUG 31	\$ Park Restaurant	Restaurants		-\$7.75
<input type="checkbox"/>	AUG 31	\$ Wal-Mart	Groceries		-\$15.39
<input type="checkbox"/>	AUG 31	ATT Cell Phone	Mobile Phone		-\$27.10
<input type="checkbox"/>	AUG 31	Aldi	Groceries		-\$0.10
<input type="checkbox"/>	AUG 31	Aldi	Groceries		-\$51.19
<input type="checkbox"/>	AUG 30	Kohl's	Clothing		-\$71.36
<input type="checkbox"/>	AUG 29	DC Parking Meters	DC Vacation		-\$3.00

Identify the **Who**: _____

Identify the **What**:

- Once we have identified the variables, we notice that some variables are numbers, like _____ and some variables are words, like _____.

- _____ variables must be numbers **AND** you must be able to perform mathematical operations on the values.
- Ask yourself this question: “**If I took the average, would the result make sense?**”
- Does it make sense to compute the average transaction amount? _____, so “**Amount**” is quantitative.

- _____ variables are usually words (but they could be numbers). These variables classify individuals.
- “**Category**” in the Mint.com example is categorical. Some transactions are “**Groceries**”, others are “**Clothing**”, and so on.

Note: There are different statistical methods for dealing with quantitative or categorical variables. Always be able to identify the type of variable you’re dealing with.

Example: Here is a data table, but only four rows are shown – the original data set had 1728 rows. Homes that were sold a few years ago in upstate New York were recorded by a real estate agency to try to predict selling price based on a house’s characteristics.

Identify each variable as **Quantitative** or **Categorical**

Price	Lot Size	Waterfront	Age	Heat Type	Living Area	Bedrooms	Neighborhood	Bath-rooms	Central Air
\$ 132,500	0.09	No	42	Oil	906	2	6	1	No
\$ 181,115	0.92	No	0	Elec	1953	3	5	2.5	No
\$ 109,000	0.19	No	133	Elec	1944	4	1	1	Yes
\$ 155,000	0.41	No	13	Gas	1944	3	1	1.5	No

Notes:

Identifier Variables

- Some variables merely identify, like your student ID number at Cecil College. **Don't** calculate the mean, **don't** make a graph, **don't** try to analyze it.

Housing Example: Homes listed for sale have an MLS (Multiple Listing Service) number so that realtors can access the property information across computer systems. This variable really can't be analyzed statistically.

ESPN.com Example: Were there any identifier variables in the dataset? If so, which?

Mint.com Example: Were there any identifier variables in this dataset? Which ones?

Example: In our Stat II course, students ran internet speed tests online using three different web browsers. “*Ping*”, “*Download*” speed, and “*Upload*” speed were each measured and recorded for the twenty computers in the classroom.

Row	Browser	Computer	Row	Ping	Download	Upload
1	Internet Exp	1	1	5	2.22	0.21
2	Firefox	1	1	26	1.19	29.8
3	Chrome	1	1	11	3.9	14.28
4	Internet Exp	2	1	10	1.44	31.25
5	Firefox	2	1	12	2	0.21
6	Chrome	2	1	13	20.67	29.02
7	Internet Fxr	3	1	5	1 15	0 21

Which variables are quantitative, which are categorical, and which are identifiers?

Quantitative: _____

Categorical: _____

Identifiers: _____

Lesson 2: Graphing Data

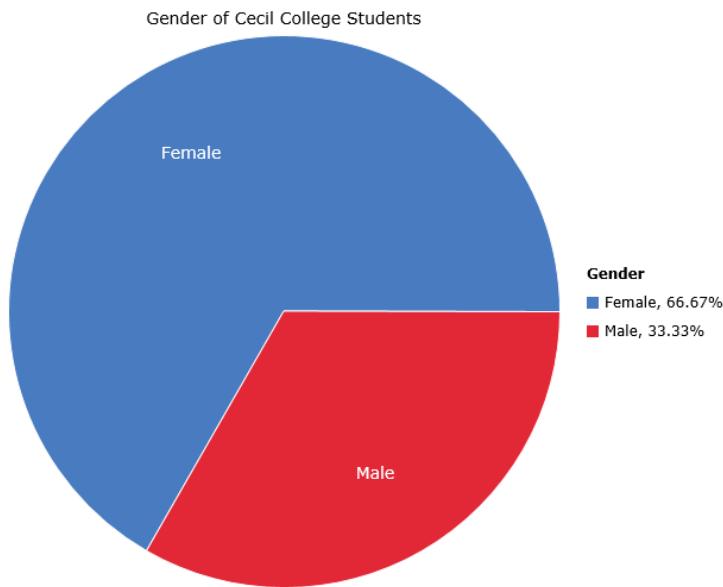
Blackboard Video 2-A

“The Lady With a Data Visualization”

1. Who was the subject of this video?
 2. What were two of her occupations?
 3. What was her conclusion about the cause of death in the Crimean War?
 4. What famous graph did she produce to prove her point?
 5. Give the **Who** and the **What** for the data that she collected. Label the variables as categorical, quantitative, or identifiers.

Blackboard Video 2-B

- When graphing data, we must first pay attention to whether the variable is categorical or quantitative.
- When graphing a **categorical variable**, your choices are _____ and _____.
- When graphing a **quantitative variable**, your choices are _____, _____, and _____.
- **All** graphs must follow the _____ principle. For example, at Cecil College, suppose 2/3 of the students are female. If we made a graph to visualize the female / male distribution, then 2/3 of the area must be devoted to the females.



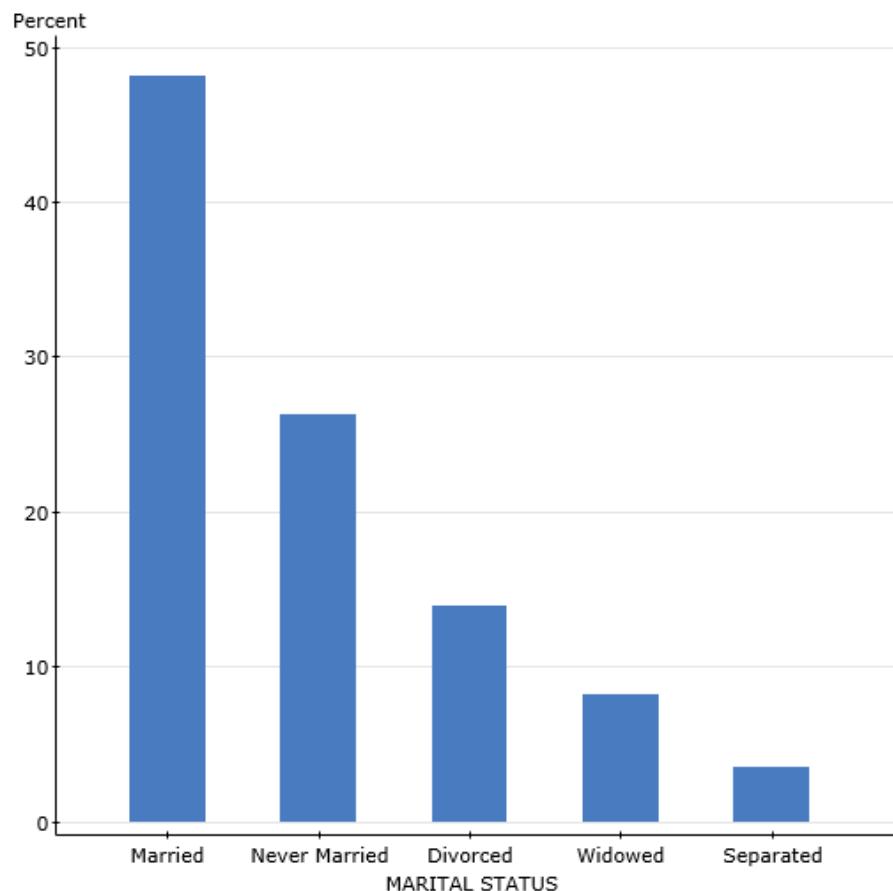
- To avoid violating the area principle, **avoid** 3-dimensional graphs. Microsoft Excel has a lot of 3D graph options – we recommend against them.
- In Math 127, we will use **StatCrunch** to make all graphs. In class, occasionally we will draw a graph by hand, but anything turned in must be created on StatCrunch.
- All graphs must include informative _____ and the axes (when appropriate) must be labeled.
- **Caution: On StatCrunch, there are no default titles, so you must add them every time.**

Graphing Categorical Variables

Example: On StatCrunch (we will access it soon), there is a dataset on the group page called the “*General Social Survey 2008*”. Researchers surveyed 2023 Americans and asked everything from “*Race*” and “*Marital Status*” to “*Do You Favor the Death Penalty*” and “*How Religious Are You*”.

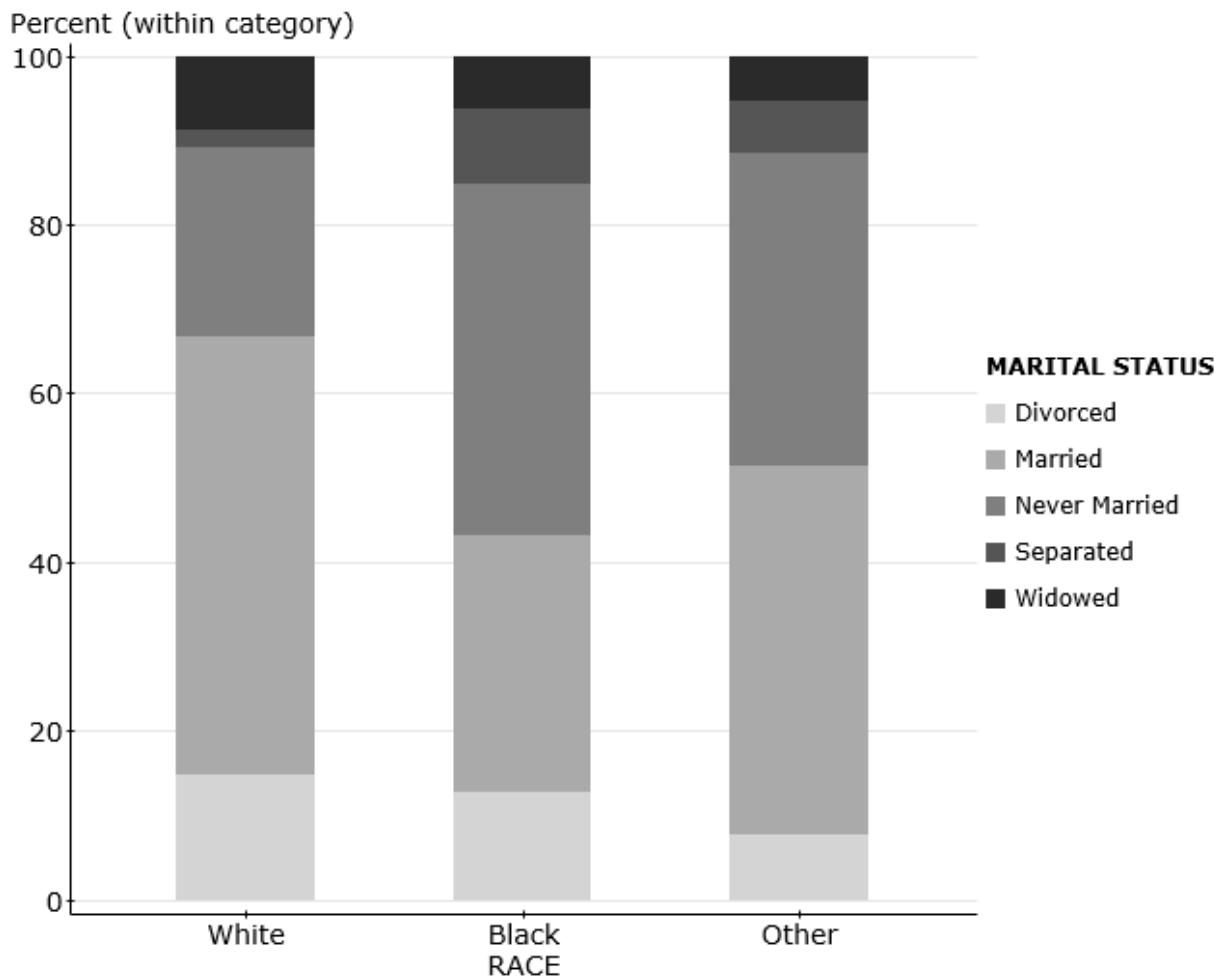
In the dataset, there were over 60 variables, some were categorical and some were quantitative.

- a. Below is a **bar plot** “*Marital Status*”. No one falls into more than one category. Approximately what percentage of respondents were divorced or separated?



- b. Give your best estimate for the number of people in the study who were divorced.

- c. Below is a **stacked bar plot**. We are looking at the distribution of “*Marital Status*” broken down by “*Race*”. What percentage of people are “*Married*” for each “*Race*”?

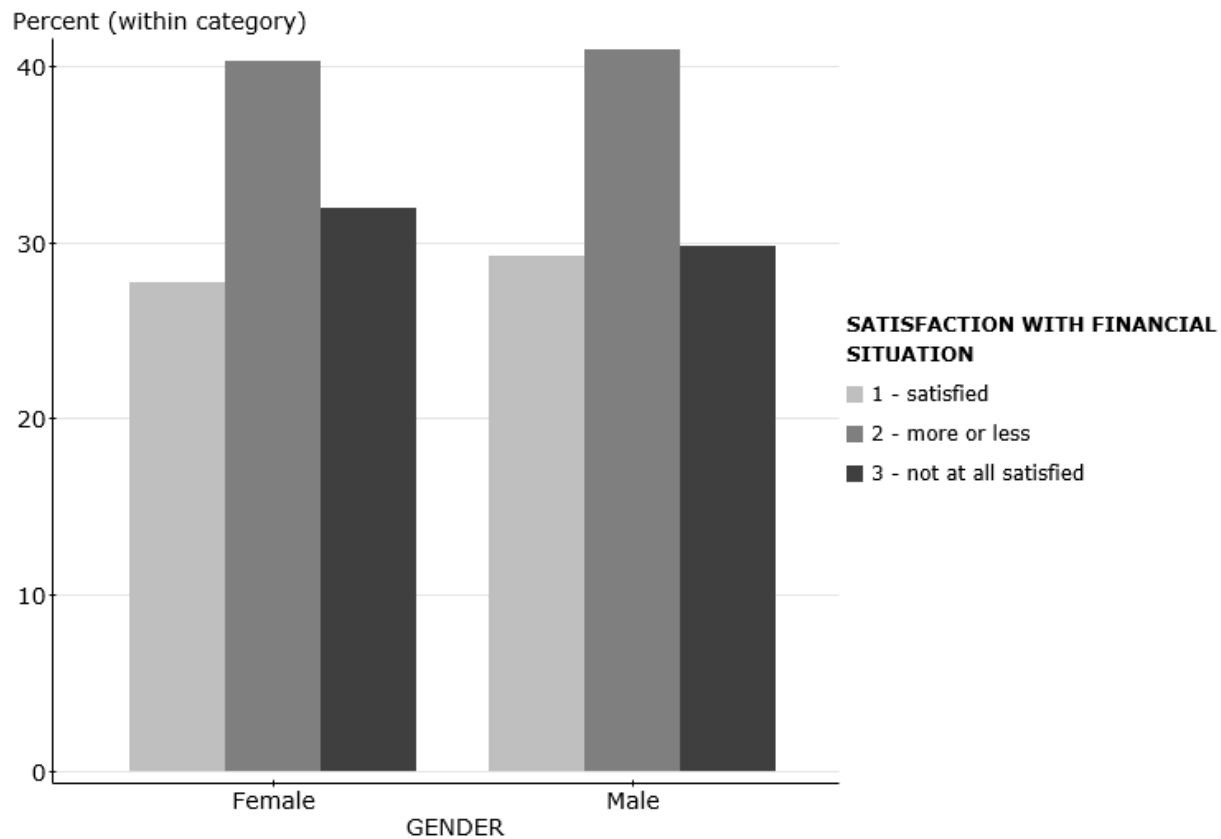


- d. Two variables are **independent** of each other if one has no impact on the other. Is “*Race*” independent of “*Marital Status*”? Why or why not?

Blackboard Video 2-C

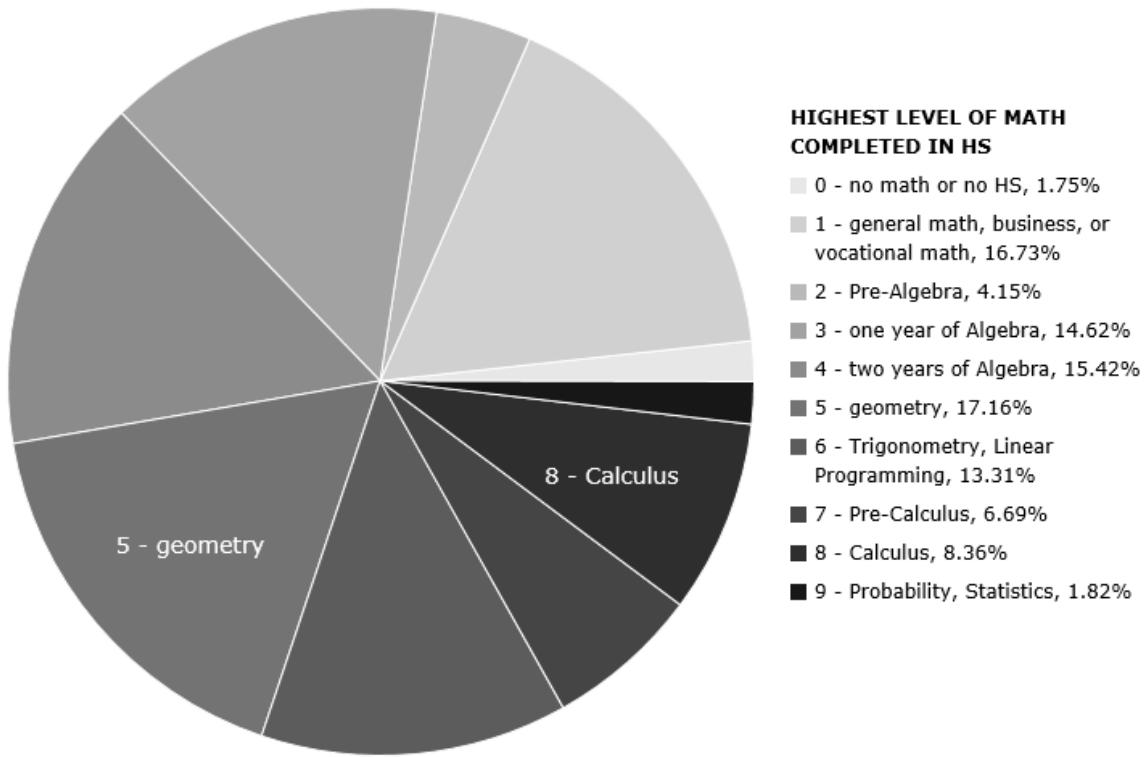
- e. Below is a **split bar plot**. We are looking at the distribution of “**How Satisfied Are You With Your Financial Situation**” across the two genders.

Based on the graph, would you conclude that financial satisfaction and gender are independent or dependent? Why?



- f. Add up the percentages for each gender. What do you get each time?

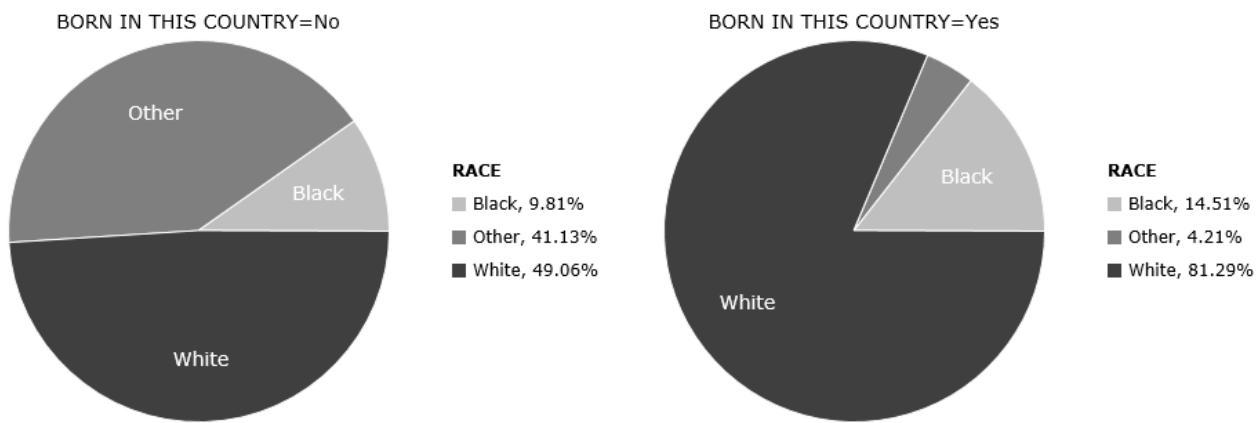
- As long as there are not too many categories, a **pie chart** makes a fine display of a categorical variable.
- g. A **pie chart** was created to analyze “**Highest Level of High School Math**”. This graph does not help us to understand the distribution. A **bar plot** or a **table** (Lesson 3) would be a better choice.



- h. What percentage of respondents never had probability or statistics?

- i. **Side-by-side pie charts** were created to look for differences in “**Race**” depending on whether or not a person was born in the USA.

“**Race**” and “**Birth Country**” are **dependent**. Why?



Bar Plots and Pie Charts – Wrap Up

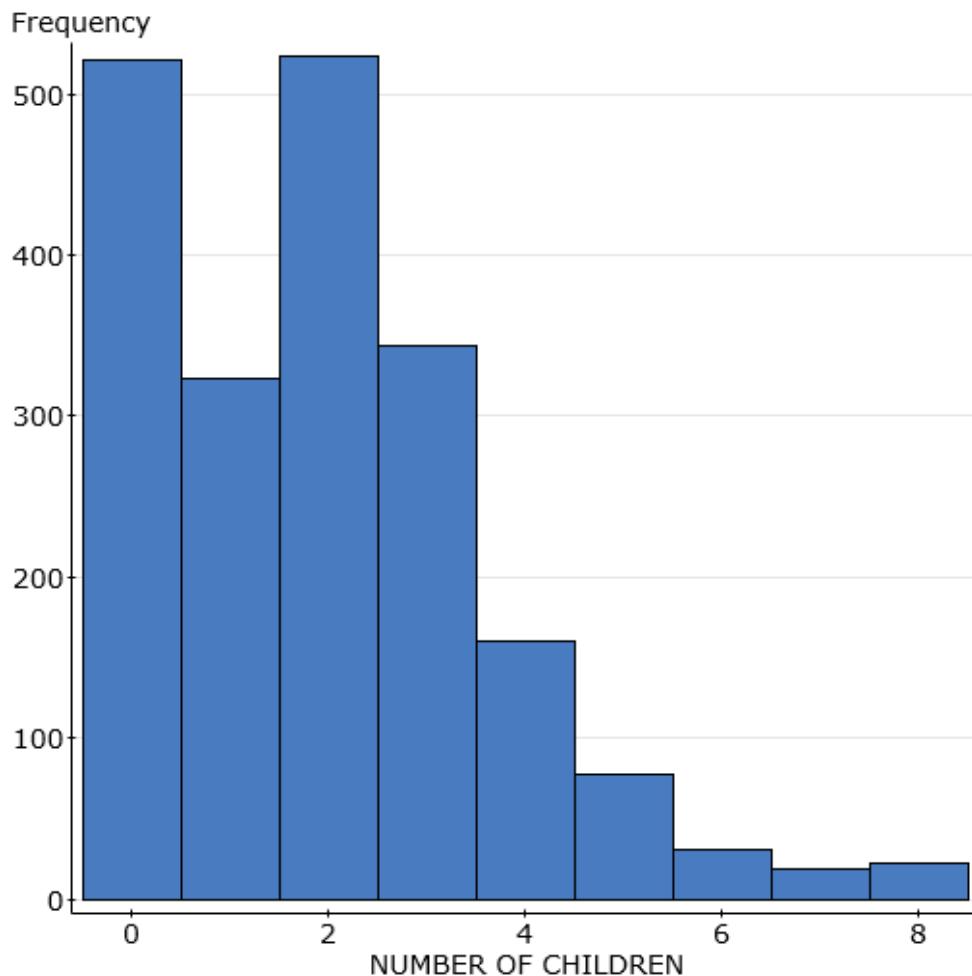
- Often one graph looks better than the other. Make both and pick the one that best tells the _____ of the data.
- Because both graphs say the same thing, there is no need to produce both.
- When we get to producing these graphs on StatCrunch during class, play with the options to make an eye-pleasing display. There is a bit of artistic flexibility, but be sure to have clear, uncluttered displays that are well-labeled.

Blackboard Video 2-D

Graphing Quantitative Variables

- When graphing a **quantitative** variable, start by making a _____. This is the primary and most important graph you can make for a quantitative variable.
- Dotplots** are best for StatCrunch, **Boxplots** come later, **Histograms** for now.

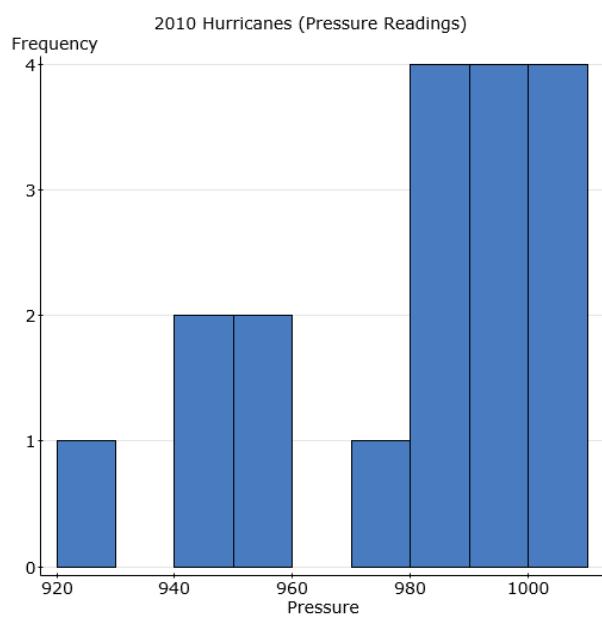
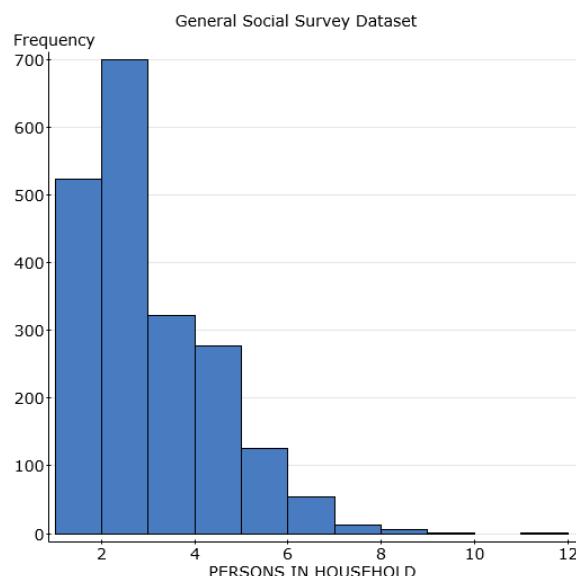
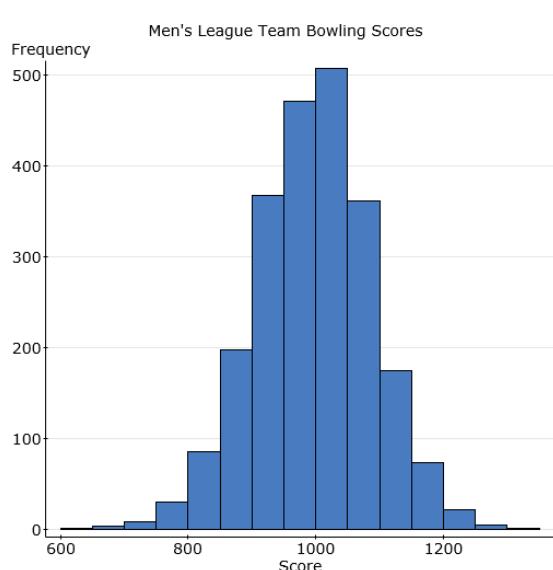
Example: Continuing with the “*General Social Survey 2008*” dataset, we explore the distribution of “*Number of Children*”.



- From this graph, we must take note of the _____, we can estimate the value of the _____, we can estimate the _____ by looking at the range of values, and we can list any unusual features or outlying values.

Shape

- If the histogram is roughly a mirror-image if you were to fold the graph in half, we say the data are _____.
- If the long tail of the distribution is to the right, we say the data are _____.
_____. **This means most of the data are piled up on the left!**
- If the long tail of the distribution is to the left, we say the data are _____.
_____. **This means most of the data are piled up on the right!**



- With shape, we also must note the number of _____.
- If a variable has **one major peak**, we call it _____.
- Two major peaks**, _____. Could even be _____.
- If the bars in a histogram are all roughly the same height, there is no mode and the data are called _____ (would be symmetric too by default).

Example: Return to the previous two pages and note the shape (skewness and modality).

Example: Think about what the **shape** of data would look like for the following variables. Draw a smooth curve to represent the distribution of the variables. Label the x -axis. Estimate the center.

a. **Home Values in Cecil County**

b. **Final Numerical Grades in Math 127 Last Semester**

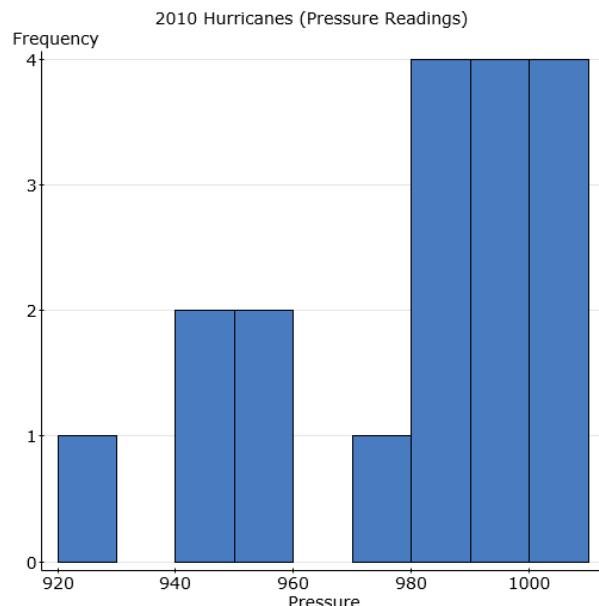
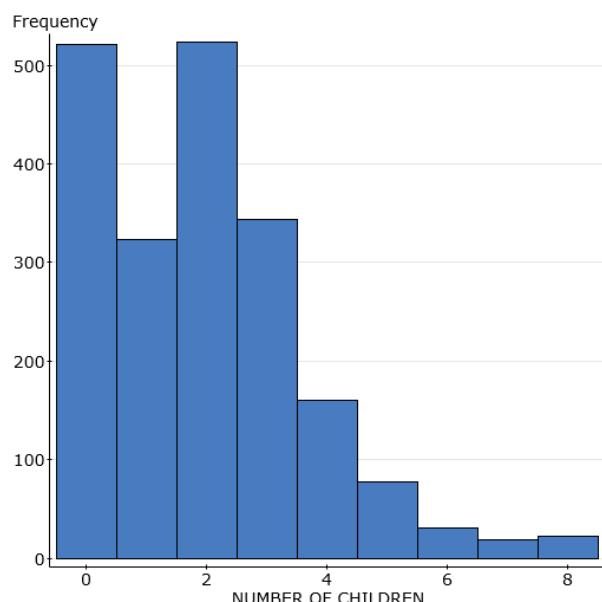
Blackboard Video 2-E

Estimating Center, Spread, and Unusual Features from Histograms

Center

- We need to eyeball a single value that accurately describes the center of the distribution. We'll do better once we start calculating summary statistics in Lesson 4, but for now, pick the value that splits the area in the graph 50/50. That is an estimate of the _____.

Example: Estimate the median value for each of the following histograms.



Spread

- We need to report a range of values that covers most, if not all of the data. Outliers can be a little tricky to deal with.

Example: Discuss the spread for the two variables pictured above.

Unusual Features

- If the variable has an unusual shape, note it down.
 - If there are unusually low or unusually high values, **note them down specifically**.
 - Don't be lazy and say, "*There are outliers*", "*There are unusual features*", etc.... Give the details!

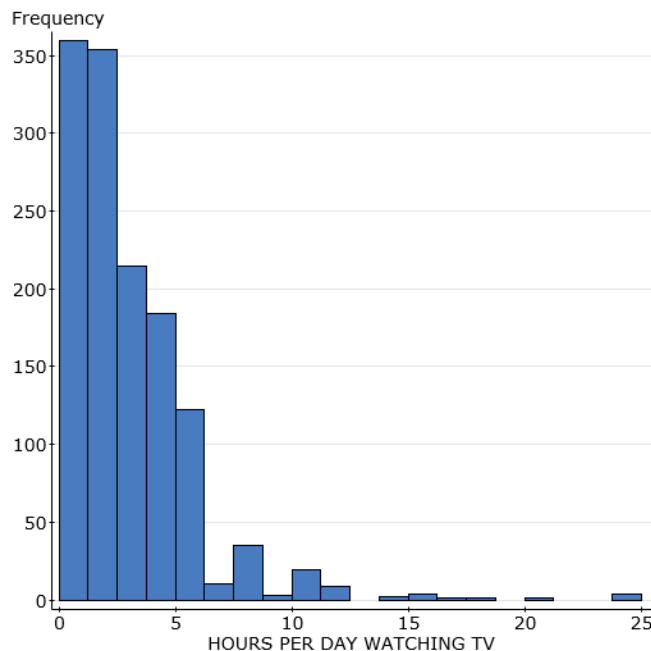
Describing a Quantitative Variable

You will be asked to describe the distribution of a quantitative variable. You should mention all features about the _____, estimate the _____, describe the _____, and give details about any _____.

If there are **no unusual features**, say so.

Write in **complete sentences**, because this is college.

Example: From the “*General Social Survey 2008*” dataset, we have “*Hours Per Day Watching TV*”. Describe the distribution of the variable.



Lesson 3: Summarizing Categorical Variables

Blackboard Video 3-A

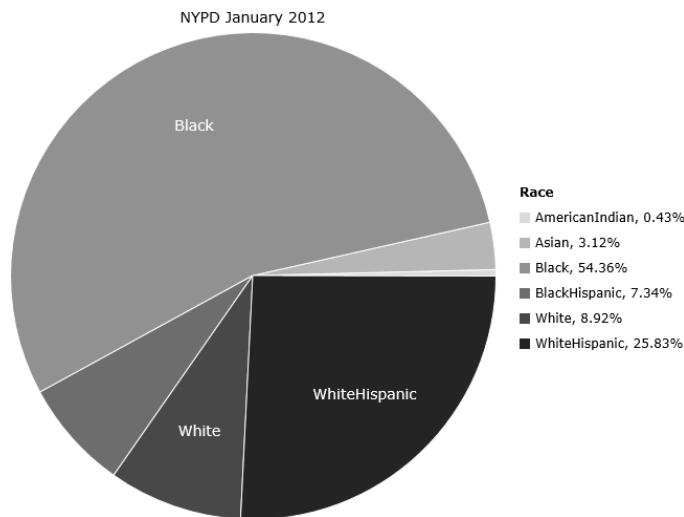
- Categorical variables, in their raw state, are usually just lists of words.
- We can summarize this data by _____, though usually we will let StatCrunch handle those gritty details.
- The primary summary statistic for a **categorical variable** is a _____ (decimal form) or a _____ (add the % sign, move the decimal two places to the right).

Example: Below is a small portion of records from the New York City Police Department from January 2012. Every person pulled over, questioned, stopped on the street, etc... is included. Believe it or not, there were 69,073 police interactions in New York City during that one month.

Row	Gender	Race	Frisked	Found Something	Arrest Made	
1	Male	Black	Yes	No	No	
2	Male	WhiteHispanic	No	No	No	
3	Male	Black	No	No	No	
4	Male	Black	No	No	No	
5	Male	WhiteHispanic	No	No	No	
6	Female	Black	Yes	Yes	Yes	
7	Male	WhiteHispanic	Yes	No	No	
8	Male	Black	Yes	No	No	
9	Male	Black	Yes	No	No	

- Each variable in the dataset is categorical, so the graphs we could make include (a few appear on the next few pages):
- When looking at **one variable at a time**, we can make a _____ table to explore the proportions and percentages.
- When looking at how **two variables relate to each other**, we can make a _____ table.

- Here is a pie chart of “***Race***” with an accompanying frequency table:



- What percentage of police questionings involved people of Hispanic descent?
- What percentage of questionings involved people that were not White?
- According to Census figures, 25.1% of the population in NYC classified as Black, yet 54.4% of the police interactions involved Black folks. Would you say that “***Race***” and “***Propensity to be Questioned by the Police***” are independent or dependent? Why?

- Now let's explore the relationship between “***Gender***” and the likelihood of getting “***Arrested***” if you’re being questioned by the police. Since we’re looking at **two variables**, here is a **contingency table**:

	No	Yes	Total
Female	4498	367	4865
Male	60759	3449	64208
Total	65257	3816	69073

d. What percentage of people stopped were arrested?

e. What percentage of females stopped were arrested?

f. What percentage of males stopped were arrested?

g. What percentage of people stopped were males?

Typo in Video – Correct Answer: $64,208 / 69,073 = 0.9296$

h. What percentage of people arrested were males?

Confusion in Video – As written in notes, answer is $3449 / 3816 = 0.9038$

Blackboard Video 3-B

Example: The table given compares what a sample of Cecil County High School students did after graduation in 1990, 2000 and 2010.

	1990	2000	2010
Continuing Education	32	60	84
Employed	41	49	25
In the Military	17	26	17
Other	10	15	14

- a. Give the marginal distribution for what graduates did after high school.

- b. What percent of 1990 graduates joined the military?

- c. What percent of 2010 graduates went on to college?

d. Give the **conditional distributions** for post graduation activities for 1990, 2000, and 2010.

e. Using your conditional distributions from part d, write a sentence or two describing the trends in the data.

f. Is year of graduation independent of post-graduation plans? Why?

Blackboard Video 3-C

Example: Back in 2003, a magazine reported on the Houston Independent School District's magnet school program.

- Of the 1755 qualified applicants, 931 were accepted, 298 were waitlisted, and the rest were turned away for lack of space.
- Of the accepted applicants, 305 were Black / Hispanic, and 292 were Asian.
- In total, there were 942 white applicants, of which 150 were waitlisted.
- Altogether, there were 373 Asians, and 50 Asians rejected.

a. Fill in the contingency table below.

	Accept	Waitlist	Reject	Total
White				
Black / Hispanic				
Asian				
Total				

b. Give the marginal distribution for “*School Acceptance*” status.

c. Give the three conditional proportions for “***Being Accepted***”, based on “***Race***”.

d. Are “***Race***” and “***School Acceptance***” independent or dependent? Why?

e. What percentage of applicants were “***White***” and “***Accepted***”?

f. What percentage of applicants were “***Asian***” or “***Waitlisted***”?

Typo in Video – Correct Answer: $640 / 1755 = 0.3647$

Blackboard Video 3-D

Independence Versus Dependence

Definition: Two variables are said to be **independent** if the values of one variable do not affect the values of the other variable. **Independent variables are unrelated.**

Example: Let's say for Cecil College students, the variables are "**Gender**" and "**Age**" and in all likelihood, these variables are independent of each other.

- We know that roughly two-thirds of students at Cecil are female, leaving one-third males.
- a. If we restrict ourselves to looking at just the 20-year-old students, what percentage will be female?
- b. If we look only at the population of 35-year-old students, what proportion do we expect to be male?

Definition: Two variables are said to be **dependent** if the values of one variable affect the values of the other variable. **Dependent variables are related in some way.**

Example: At Cecil College, "**Gender**" and "**Are You An Engineering Major**" are definitely **dependent** variables.

While only **one-third** of our students are males, if you walk into any Calculus III, Linear Algebra, or Engineering course, you are likely to see **80%** or more of the students in these classes as males.

The same would hold true for "**Are You A Nursing Major**", except in the opposite direction. Typically females hold over 90% of those spots.

We'd have to collect data to analyze "**Are You A Business Major**". That one might be pretty close to being independent.

Showing Variables Are Dependent When Using a Contingency Table

1. Compute the appropriate **marginal proportions** by taking row totals (or column totals) over the grand total. If you did this for every row (or every column), you'd have the **marginal distribution**.
2. If the marginal distribution computed in Part 1 was for the row variable, create the same distribution, except look only at one column at a time. This is a **conditional distribution**.
3. If the row variable and the column variable are **dependent**, then your marginal distribution and conditional distributions will **differ** substantially. This is more likely.
4. If the row variable is **independent** of the column variable, **every conditional distribution will be equal to the marginal distribution**. This is less likely.

Note: Almost all variables are dependent to some degree, so aim to show this by finding at least one substantial difference from the marginal distribution to the conditional distributions.

Example: In our NYPD dataset, below is a contingency table to analyze “***Race***” versus “***Frisked***”.

	American Indian	Asian	Black	Black Hispanic	White	White Hispanic	Total
No	158	1156	15505	1932	3420	7392	29563
Yes	129	945	21043	3000	2577	9978	37672
Total	287	2101	36548	4932	5997	17370	67235

- a. Give the marginal distribution for whether or not a person was frisked.

b. Give the conditional distribution of being frisked for each “**Race**”.

c. Are “**Race**” and “**Being Frisked**” independent or dependent? Why?

Lesson 4: Summarizing Quantitative Variables

Blackboard Video 4-A

Hans Rosling Video: “Above Average”

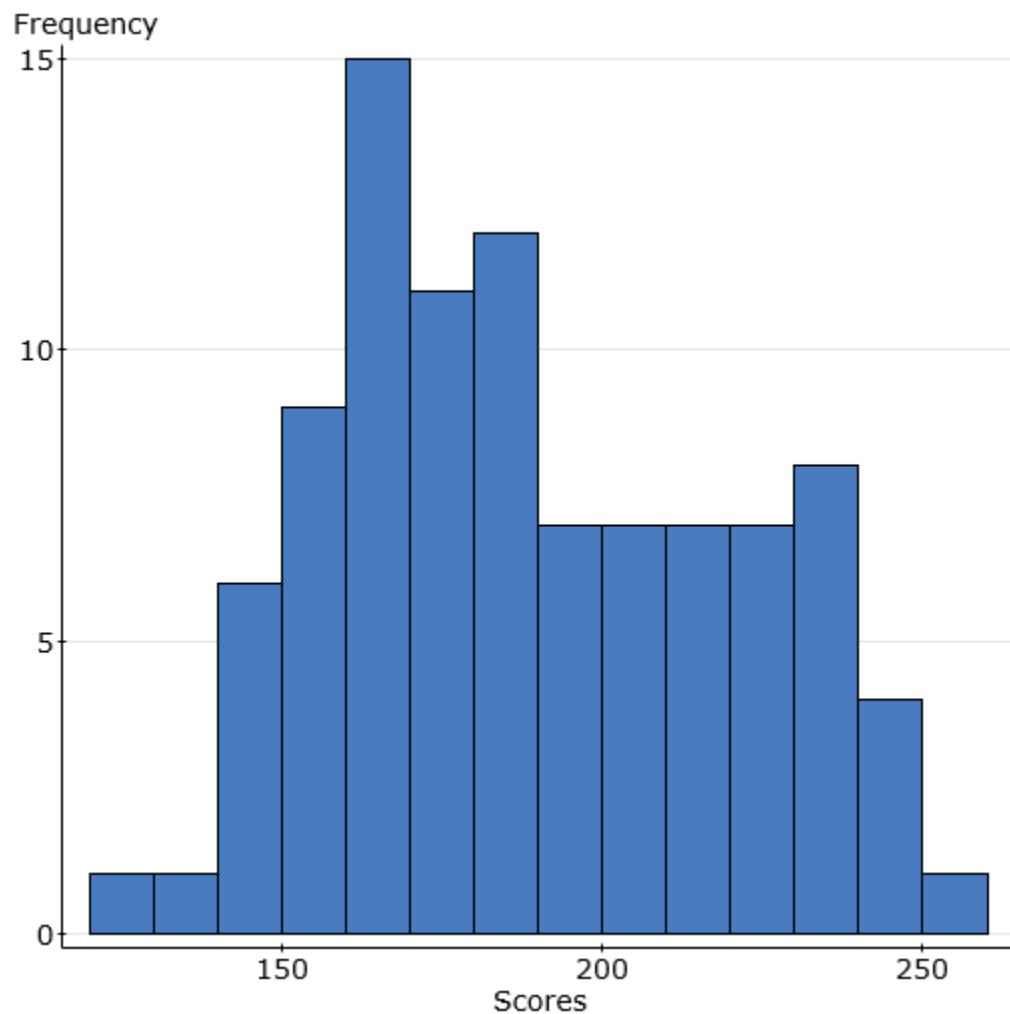
Blackboard Video 4-B

- To summarize a **quantitative variable** with the appropriate summary statistics, we must first start by looking at a histogram to determine if the shape is roughly symmetric or skewed.
- When describing a quantitative variable, we must mention _____, _____, _____, and give details about any _____.
- No longer will we eyeball the center and spread. We will calculate the best statistics and report those values.

Describing the Center of a Distribution: We have two main options for a single number that describes the center of a distribution, the _____ and the _____.

- The _____ is best for summarizing symmetric data. Why?
- The _____ is best for summarizing skewed data. Why?

Example: Here are Professor Kupe's bowling scores from a recent Men's League season. This data can be characterized as skewed right (not severely, but it isn't symmetric).



- Which measure of center will be higher, his mean or his median bowling score? Why?
- Interpret the value of the median with a sentence in context.

Calculating the Mean By Hand

Example: An article reported the single-leg power measurements for a sample of 14 elite endurance cyclists.

244	191	160	187	180	176	174
205	211	183	211	180	194	200

Calculate the mean by hand:

Calculating the Median by Hand

Example: For the cycling data on the previous page, calculate the median by hand.

What would happen to the mean and the median if the guy with leg strength of 244 actually had leg strength of 205?

- The median is _____ to outliers. That's why we use the median if the data are skewed.
- Often times, very skewed data like salary, home value, net worth, or grades in a college course will have the median reported rather than the mean.
- In Math 127, we typically will calculate our summary statistics using StatCrunch. We prefer to demonstrate the ideas with the formulas to help our understanding. StatCrunch will be emphasized during class meetings.

Blackboard Video 4-C

Describing Positions in a Dataset Other than the Center

Example: Below are cycling data from the previous example, sorted for our convenience.

160	174	176	180	180	183	187
191	194	200	205	211	211	244

- The _____ is the value that splits the dataset into two pieces, a bottom half and a top half.
- We can also split the dataset into four quarters. The values that do this are called the _____.
- _____ is the value that splits the bottom 25% from the top 75%.
- _____ is the value that splits the bottom 50% from the top 50%. It's also the median.
- _____ is the value that splits the bottom 75% from the top 25%.
- Typically StatCrunch will calculate quartiles for us, so we won't bother with a formula. We can use a bit of common sense to figure them out if we need to.

Sketch out a generic diagram of the quartiles:

Determine the quartiles for the cycling data above:

- We can also chop up a dataset into 100 chunks. The values that do this are called _____.
- Statisticians cannot agree how to calculate percentiles, so we won't bother with a formula.

Sketch out a generic diagram of the percentiles:

Example: When you took the SAT, suppose you scored in the 22nd percentile on your math score. What does this mean?

Example: Suppose you also scored at the 3rd quartile for your English score? Interpret.

Example: Professor Kupe can remember working with a guy who's wife just had their first baby. He proudly mentioned to me that the baby's skull circumference was in the 98th percentile. What's that mean?

Blackboard Video 4-D

Describing Spread

Knowing the center and other measures of position of a distribution doesn't tell us the whole story – as important or *even more important* is coming up with ways to characterize the amount of spread in a distribution.

Range

- The range = _____ is a single number that tells us the difference between the extremes.
- It uses only _____ data values so it has limited usefulness.

Interquartile Range

- Another measure of spread, the _____, looks at the range of the middle 50% of the data. To calculate this, we need to get the quartiles for the dataset first. Then find the difference between the first and third quartiles:

Example: Calculate the range and IQR for the cycling data.

160	174	176	180	180	183	187
191	194	200	205	211	211	244

- How small can IQR get?

IQR versus Standard Deviation

- For any distribution, the **IQR** is a reasonable measure of spread, but it has some drawbacks.
- For one, the IQR only uses limited information in its computation. We actually discard the bottom 25% and top 25% of the data.
- The IQR is best for data that is _____.
- If there are a few outliers, the IQR won't be affected. For this reason, the IQR is best for skewed data.
- For skewed data, use the _____ to describe the center and IQR to describe the _____.
- If data is symmetric, we can do better than the IQR.
- For symmetric distributions, use the _____ to describe the center and the _____ to describe the spread.
- Symmetric data typically won't have outliers piled up on one end of the distribution. Outliers distort the mean and _____ the standard deviation.

Common Language (Real World) Definition of Standard Deviation:

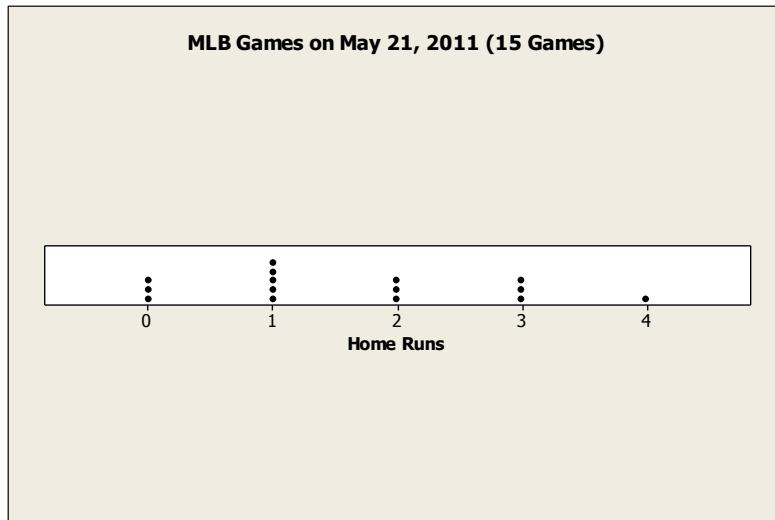
Formula Definition for Standard Deviation:

Blackboard Video 4-E

Example: Compute the standard deviation by hand for the cycling data.

160	174	176	180	180	183	187
191	194	200	205	211	211	244

Example: If we pick Saturday May 21, 2011 as a representative day and take every Major League Baseball game from that day as our sample, below is a dotplot of the number of home runs from each game.



- a. Recreate the dataset from the dotplot.

- b. Calculate the mean and standard deviation by hand.

- c. Refer back to the dotplot and visualize how the standard deviation is “like” the typical or average distance to the mean.

Blackboard Video 4-F

Example: Cook up a dataset with each of the following characteristics. There are an infinite number of answers for each part.

a. **8 Values:** Mean way bigger than the standard deviation.

b. **6 Values:** Standard deviation way bigger than the mean.

c. **5 Values:** Standard deviation equal to 0.

d. **9 Values:** IQR equal to 0 and standard deviation bigger than 10.

e. **10 Values:** Mean is negative, Median = Q_3 = Maximum are all positive.

Blackboard Video 4-G

We turn our attention to comparing groups.

- We can compare **histograms** to compare and contrast shape, center, spread, and unusual features.
- We can compare **boxplots** to see how the **5-number summaries** differ.

5-Number Summary: _____

Example: During last summer at Cecil College, Math 127 students reported their age and gender. Determine the **5-number summary** for each gender; the data have been sorted for our convenience.

Females

18	18	18	19	19	19	19	19
19	19	19	20	20	20	21	21
23	26	29	32	33	40	45	47

Males

19	19	19	19	19	20	21	22	27	56
----	----	----	----	----	----	----	----	----	----

The Outlier Rule-of-Thumb

- We would like to officially classify extremely high or low values as outliers. Here is the rule we will use:

Example: Determine if there are any official outliers in the Cecil College age dataset from the previous page.

Blackboard Video 4-H

Drawing a Boxplot (StatCrunch Will Do This)

- Boxplots are great graphs to compare the major measures of position among multiple datasets. Here are the steps:
 1. First calculate the 5-number summary and determine outliers using the rule of thumb.
 2. Place lines at the three quartiles and connect them with a box.
 3. Draw “whiskers” to the most extreme actual data values that are contained inside the upper and lower fences.
 4. Mark any official outliers with a dot or asterisk.

Example: Draw side-by-side boxplots to compare ages of Cecil College Math 127 students.



Subtle Note: Histograms and dotplots are graphs of **data**. Boxplots are graphs of **5 measures of position**. Be careful when describing the shape based only on a boxplot – you should always look at a histogram when possible.

Dealing with Outliers

- Many times, the outliers in a data set are going to be the most interesting cases, but dealing with them is a delicate matter.
- The outlier rule-of-thumb identifies outliers, but it doesn't tell you what to do with them.
- _____ silently leave an outlier in place and pretend as if nothing is unusual.
- _____ drop an outlier from the analysis without comment just because it is unusual.
- If an outlier is simply unbelievable, perhaps from an incorrect response, transposing numbers, etc..., then try to **investigate** and correct the **number**.
- If there is no way to correct the value, many times the best path is to **run the analysis both with and without the outlier**.

Example: In 1961 Roger Maris made baseball history by hitting 61 homeruns, eclipsing Babe Ruth's 1927 record by one. Here are Maris's homerun totals for his 10 seasons. Is his record-setting year considered an outlier? Explain. What does common sense tell us? Draw a boxplot.

8	13	14	16	23	26	28	33	39	61
---	----	----	----	----	----	----	----	----	----

Blackboard Video 4-I

- At times, we will take our data and change it by performing a mathematical function on every data value.

Example:

- In your sociology course, your instructor awards a 3-point curve on the final exam – every final exam score now has 3 points added to it. What happens to the summary statistics?
- Volunteers in a weight-loss study have their net weight loss measured in pounds. Before publishing the results, the researcher convert pounds to kilograms by multiplying every data value by 0.4536. What happens to the summary statistics?

Shifting and Rescaling

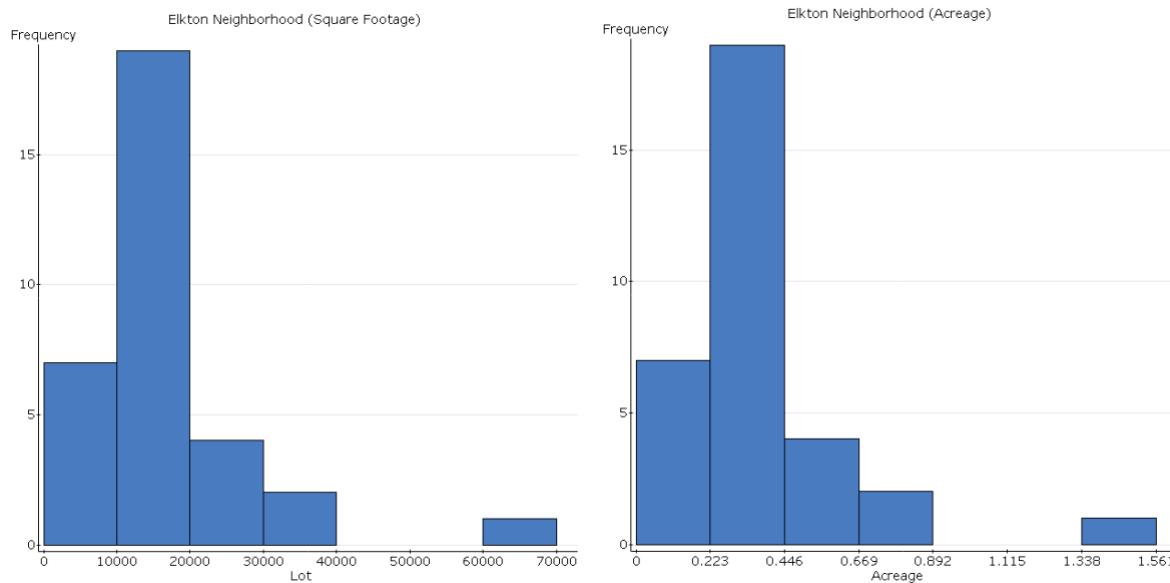
- When we **shift** data by **adding** or **subtracting** a constant to every data value, all measures of _____ will be changed by that amount.
Measures of spread will remain _____.
- When we **scale** data by **multiplying** or **dividing** by a constant, **all** measures of position and **all** measures of spread will be _____.

Make a List of the Measures of Position:

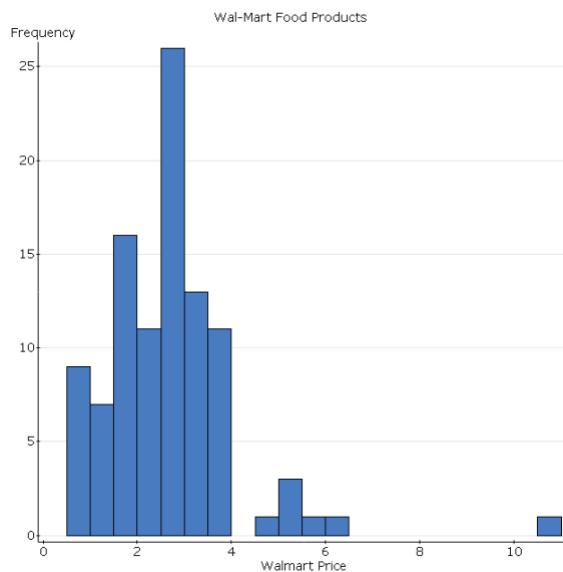
Make a List of the Measures of Spread:

Example: In one neighborhood in Elkton, lot sizes are listed in square feet. The mean lot size is 15,746 with a median of 11,282, a minimum of 6,664 and a maximum of 64,904. The standard deviation was 11,511 and the IQR was 6,796.

- a. Do you think the distribution is skewed or symmetric? If skewed, which way? Why?
 - b. If lot sizes were expressed in acres rather than square feet, what would the new summary statistics be? There are 43,560 square feet to 1 acre.
 - c. Notice the similarities and differences between the original data and the scaled data.



Example: Cecil students went to Wal-Mart a few semesters ago and collected pricing data on food products. The histogram and summary statistics are listed below.



Column	n	Mean	Std. dev.	Median	Range	Min	Max	Q1	Q3	IQR
Walmart Price	100	2.70	1.40	2.58	10.29	0.69	10.98	1.81	3.29	1.48

- a. If Wal-Mart ran a “50 cents off” every food item special, what would happen to each statistic above?

Column	n	Mean	Std. dev.	Median	Range	Min	Max	Q1	Q3	IQR
Walmart Price	100									

- b. If Wal-Mart ran a “10% off” every food item special, what would happen to each statistic above?

Column	n	Mean	Std. dev.	Median	Range	Min	Max	Q1	Q3	IQR
Walmart Price	100									

Blackboard Video 4-J

Z-Scores

Main Idea: When we want to determine how unusual a data value is, usually this is done
_____ and _____.

Example:

- You are in Levi's English 101 and your final grade was an 88 and in Vinton's Biology 208 your final grade was an 81.
- Your boyfriend (girlfriend) is in the same classes – he (she) got an 80 in English but a 90 in biology. He's claiming he did better than you during the semester since Biology is harder and if you add up his final grades, the sum is a bit higher than your sum.

Can we put him (her) back in his place by standardizing the grades and looking at z -scores?

Definition: A z -score is:

The z -score formula for sample data is:

Class	Your Score	His Score	Class Mean	Class Standard Deviation
English	88	80	75.5	5.1

Class	Your Score	His Score	Class Mean	Class Standard Deviation
Biology	81	90	61.0	13.2

Note:

- Observations within _____ standard deviations of the mean are for the most part considered not unusual.
- Observations between _____ and _____ standard deviations of the mean are “unusual”.
- Observations more than _____ standard deviations from the mean are “rare”.
- A z -score that is positive simply indicates that a data value is _____ the mean.
- A z -score that is negative simply indicates that a data value is _____ the mean.
- Converting a data value to a z -score is called _____.
- If asked if a data value is an outlier, use the _____.
- If asked if a data value is “unusual” or “rare”, convert to a z -score and use that to make your determination.

Example: Kiplinger's ranks colleges each year, and we have data for the “**100 Best Values in Public Colleges**”. The variable of interest for this problem is “***Debt at Graduation***”. The data were reasonably symmetric, so means and standard deviations are appropriate statistics.

Summary statistics:

Column	n	Mean	Std. dev.	Min	Q1	Q3	Max
Debt At Graduation	98	\$20,038	\$4,157	\$10,927	\$16,824	\$22,102	\$30,760

- a. Students of James Madison University should expect \$18,183 in debt upon graduating. Convert this value to a z -score. Is this an unusual amount of debt?
 - b. Give a range of debt value that would not be unusual.
 - c. A certain university had a z -score of -1.63 . Compute the debt.

Lesson 5: Collecting Data With Observational Studies or Surveys

Blackboard Video 5-A

“Battling Bad Science” with Ben Goldacre

- a. According to Goldacre, what is epidemiology?

- b. What both causes and prevents cancer?

- c. What is the absolute weakest form of evidence?

- d. “*Olive Oil Prevents Skin Wrinkles*” is an example of an **observational study**. When using this type of study, **lurking variables** are possibly the reason why the researchers make the conclusions that they do. Goldacre suggests a number of lurking variables for the olive oil study. Name two of them.

- e. According to Goldacre, in what book was the first trial ever recorded?

- f. To help with pain, what worked better than a sugar pill?

- g. Goldacre mentions trials (designed experiments, Lesson 6) and notes that one major flaw repeatedly exercised by pharmaceutical companies is testing the new drug against a _____. What should happen is the new drug should be tested against _____.
- h. According to Goldacre, which trials **almost always get the results that they want**: Industry-funded trials or independent trial?
- i. According to Goldacre, which trials use **better** methods: Industry-funded trials or independent trials?
- j. How does Goldacre reconcile the apparent mismatch in parts h. and i.? ***Hint:*** It has to do with the data collected in the industry-funded trials.
- k. Around _____ of all of the trial data on antidepressants have been withheld.

Blackboard Video 5-B

- We are always interested in the whole group, the _____, but it is usually impractical (or unnecessary) to examine everybody or everything. Summary numbers that describe a population are called _____.
- The solution is to just look at part of the whole, a _____. Summary numbers that describe a sample are called _____.

Example: At Cecil College, money was spent to refurbish the physical education building. Are students using it? Draw a picture of the situation. Suppose we are asking the question, ***“Do you regularly use the physical education building for any purpose aside from your normally scheduled classes?”***

Blackboard Video 5-C

Definition: An _____ is a type of study in which individuals are observed and certain outcomes are measured. **No** attempt is made to affect the outcome. **No** treatment is given. One might call this the “**hands off**” method of obtaining data.

Example: At the end of the semester, a professor notices that students who missed at least 3 classes performed much worse than students who missed at most 2 classes.

Example: A team of medical researchers track many pregnant women over a number of years. A highly publicized 2008 study found that women who consumed 200 mg or more of caffeine a day had double the risk of miscarriage of those who took in no caffeine.

Example: Let’s call the 1950s and 1960s the “**Golden Age**” of baseball. Let’s call the 1990s and early 2000s the “**Steroid Age**” of baseball. Discuss the details of using an observational study to determine which era had more homeruns on average per at-bat.

Limitations of Observational Studies

- Observational studies are very common, but you must be careful with your conclusions.
- Because no _____ into different treatment groups has taken place, we can only look for general relationships between variables.
- We cannot prove _____ relationships when using observational studies. The data may be _____, leading us to false conclusions. There may be _____ variables that we haven't thought about.

Example:

- With the class attendance / final grade example, can we say for sure that missing 3 classes causes your final grade to decrease? Name a possible lurking variable.
- With the caffeine / miscarriage example, can we say for sure that ingesting 200 mg of caffeine per day will cause your miscarriage rate to double? Name a possible lurking variable.
- With the steroids / home runs example, name a few possible lurking variables.

Blackboard Video 5-D

- One way to obtain an observational study is to conduct a survey or draw a random sample.

Three Main Ideas of Sampling a Population

Number One: _____

- **Pot of chili analogy:**

Number Two: _____

- It should be intuitive that a larger sample is better (almost always).
- But, we don't need to make our samples sizes a fraction of the population size, like for example, 10% of the Cecil College students.
- **What matters is only sample size** – what if we just made a 20 gallon vat of chili? How big does the sample need to be?

Number Three: _____

- Randomizing protects us from the influences of all the features of our population, even ones we haven't thought about.
- It does so by making sure that on average, the sample looks like the rest of the population.

Cecil College Example – Discuss The Anti-Random Sample:

Blackboard Video 5-E

Definitions – Link Each to the Cecil College Fitness Center Example

At Cecil College, money was spent to refurbish the physical education building. Are students using it? Suppose we are asking the question, “*Do you regularly use the physical education building for any purpose aside from your normally scheduled classes?*”

Census	A sample consisting of the entire population.
	Cecil Example:
Sampling Frame	A list of individuals from which the sample is drawn.
	Cecil Example:
Simple Random Sample	A sample drawn such that it had an equal chance of being selected when put against every other possible sample of the same size. Also, every member of the population has an equal chance of being selected.
	Cecil Example:

Stratified Sample	A sampling design in which the population is divided into sub-populations first, and then random samples are drawn from each sub-population.
	Cecil Example:
Cluster Sample	A sampling design in which entire groups are chosen at random, usually done as a matter of convenience or practicality. Groups must be heterogeneous (each group mimics population).
	Cecil Example:
Systematic Sample	A sample that is drawn by taking every k^{th} individual.
	Cecil Example:

Convenience Sample	A sample that is drawn by selecting the most readily-available data without implementing randomization or any of the above-mentioned methods.
Cecil Example:	
Multistage Sample	A sample that combines any of the above listed methods.
Cecil Example:	
Nonresponse Bias	Bias introduced into the sample when a large percentage of those sampled fail to participate.
Cecil Example:	

Undercoverage	A sampling scheme that biases since it gives part of the population less representation than it really has. Cecil Example:
Voluntary Response Bias	Bias introduced when individuals can choose on their own to participate. Cecil Example:

Blackboard Video 5-F

Example: Cecil County has 10 zip codes, of which 3 were selected by the Post Master. Inside each zip code, 5 neighborhoods were selected randomly by a computer and the Saturday mail was postponed until Monday for the month of June. Post Office officials tallied up the number of complaints by residents to determine if Saturday delivery is still a necessity.

a. Draw a diagram of the population and the sample.

b. Identify all methods of sampling used in this problem.

c. List any potential sources of bias.

Example: In the warehouse, quality control managers randomly select 8 cases of Cool Ranch Doritos from the pool of all the day's manufacturing. Every bag in the 8 cases is weighed to ensure the bag-filling machinery is working properly. Once the bags are weighed, the manager grabs one from each case, opens it, and inspects it for minimal broken chips.

- a. Draw a diagram of the population and the sample.

- b. Identify all methods of sampling used in this problem.

- c. List any potential sources of bias.

Lesson 6: Collecting Data With Designed Experiments

Blackboard Video 6-A

Discuss: The developmental math sequence at Cecil College was “redesigned” Spring 2011 in attempts to increase completion. Diagram the experiment.

Main Points

- For the most part, we need to run a _____ to collect the evidence needed to show a cause and effect relationship. Observational studies will not (in almost every case) provide us with the evidence needed to prove causality.
- By actively _____ treatments applied to the experimental units, we can build a case that a certain treatment leads to a certain outcome.
- Good experimental design requires us to _____ sources of variation other than the factor we are testing. We must think about the things that might affect the results of the experiment and control them.
- _____ allows us to equalize the effects of unknown sources of variation. We must randomize the treatments when running an experiment.
- Multiple runs of a treatment on an experimental unit is a good thing; it gives us confidence in our conclusions. We call this _____.
- At times, another variable might influence the results of our experiment, but we just don't happen to be studying its effects. We can utilize _____ to remove this added variability.
- At times, a _____ variable may be the underlying reason for the results we see. Always be aware that there are other factors we may not have even considered that are influencing the results of the experiment.

Example: Think back to the developmental math “experiment” that was run at Cecil College. Comment on the main points of this lesson.

Blackboard Video 6-B

Example: Design an experiment to test if taking steroids and human growth hormone causes a major league baseball player to hit more homeruns.

Blackboard Video 6-C

Definitions – Link each to the steroids example.

Random Assignment	The process in an experiment when it is randomly decided which treatments go to which experimental units.
Steroids Example:	
Factor	A variable with values that are controlled by the person running the experiment. Factors are randomly assigned.
Steroids Example:	
Blocking Factor	A variable controlled by the person running the experiment. Blocking factors are not randomly assigned.
Steroids Example:	
Level	The specific values the experimenter chooses for the factors.
Steroids Example:	

Treatment	The exact combination of factors and their levels that are applied to the experimental units.
Steroids Example:	
Response Variable	The variable whose values are compared across different treatments. This is the observed outcome of each trial.
Steroids Example:	
Experimental Units	Individuals or objects on which the experiment is performed. Sometimes called subjects if working with people or animals.
Steroids Example:	
Control Group	The experimental units assigned to a baseline treatment, a placebo treatment, or even no treatment.
Steroids Example:	

Statistically Significant	When the observed difference at the end of an experiment is too large to attribute solely to chance, we consider the difference to be _____.
Steroids Example:	

Blackboard Video 6-D

Example: To research the effect of wearing sunglasses when playing poker, a player ran an experiment over the course of one year. Each time he went to the casino, he flipped a coin and if it came up heads, he wore sunglasses that day. He kept track of his daily net winnings for the year and compared the mean profits with and without sunglasses. He played four times each week, twice during the week and both on Saturday and Sunday.

Because the players vary wildly between “*during the week*” and “*on the weekend*”, he noted this factor down in his records. If during the week, he wore sunglasses the first day, then he didn’t wear them the second day. Same thing for the weekend.

Diagram the experiment, labeling everything.

Example: The wife of a math professor went online and bought 4 kits of 5 caterpillars each; they're supposed to turn into butterflies. Ten were kept indoors while the other ten were left out on the deck (each group inside their respective internet-provided "habitats"). In both groups, half the caterpillars were given their internet-provided food packet, but the other half were given sugar water and leaves from the yard.

Luckily, all 20 caterpillars turned into cocoons, and all 20 cocoons turned into beautiful butterflies (on 15 of them, she didn't really follow the instructions). The variable of interest was "time until butterfly".

Diagram the experiment, labeling everything.

Math 127 Unit I Checklist²

- I can differentiate between categorical variables and quantitative variables.
- I know which graphs and summary statistics are appropriate for categorical variables.
- I know which graphs and summary statistics are appropriate for quantitative variables.
- I can work with 2 by 2 contingency tables, compute marginal and conditional proportions, and understand the concept of independence.
- I know how to describe the distribution of a quantitative variable with shape, center, spread, and unusual features.
- I can compute z -scores for a given data value and compute a data value for a given z -score.
- I can describe standard deviation, z -score, and statistical significance to someone who has never taken statistics.
- I am familiar with the principles of sampling a population, the various methods of drawing a sample, and all the terminology dealing with sampling methods.
- I am familiar with designed experiments and the related terminology, and I can draw a detailed diagram of an experiment.

² Can you check these off your list? If you can't, get help (your instructor, the math lab, a peer, online).

Lesson 7: The Difference Between Data and Models

Blackboard Video 7-A

- In practice, statisticians, scientists, academics, businesspeople, and researchers typically formulate a research question or pose a hypothesis that they wish to investigate.
- To answer their questions, they first _____ their data (Lessons 5 and 6), then they _____ the data (Lesson 2), and finally they _____ the data (Lessons 3 and 4).
- Most often, researchers must utilize only a _____ of the data contained in the whole _____. For example, if the State of Maryland was investigating the percentage of residents who might take advantage of a tax credit for installing solar power grids, the State certainly could not reasonably expect to survey all 5,000,000+ Maryland residents.
- With the data in hand, typically the next step is to create a _____ using your data.
- A model is a mathematical _____ of the pattern represented in the data.

Example: 385 Ivy League undergraduate students were recruited to take part in an IQ study. The head researcher theorized that Ivy League students are more intelligent than the typical person. In general, the average IQ score is 100. At each of the eight Ivy League schools, 50 students were randomly selected and offered \$100 to participate. Due to the incentive, there wasn't much nonresponse (15 students slept in that Saturday morning). Below are the first few rows of the dataset from StatCrunch.

Row	IQ	var2	var3	var4	var5
1	90				
2	110				
3	58				
4	106				
5	96				
6	129				
7	116				

a. What kind of sampling methodology was used? Explain.

b. Explain exactly what the “*data*” is in this example.

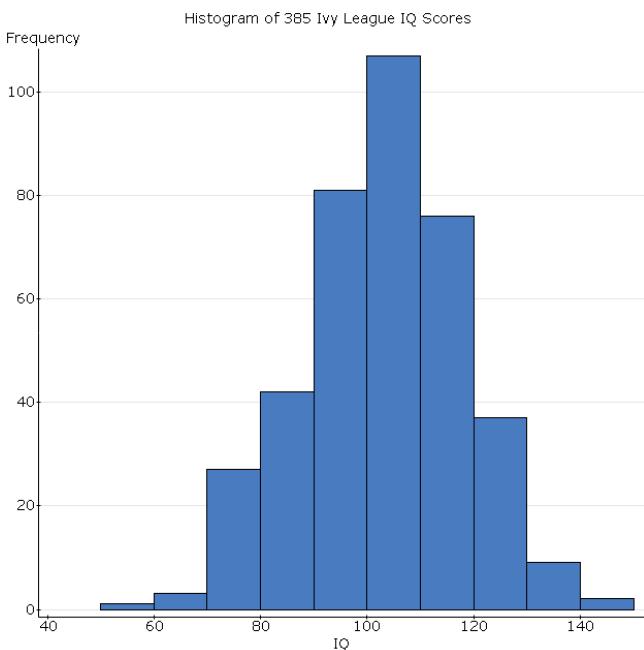
c. What is the population of interest and what is the sample?

The researchers then took the data and summarized it as follows:

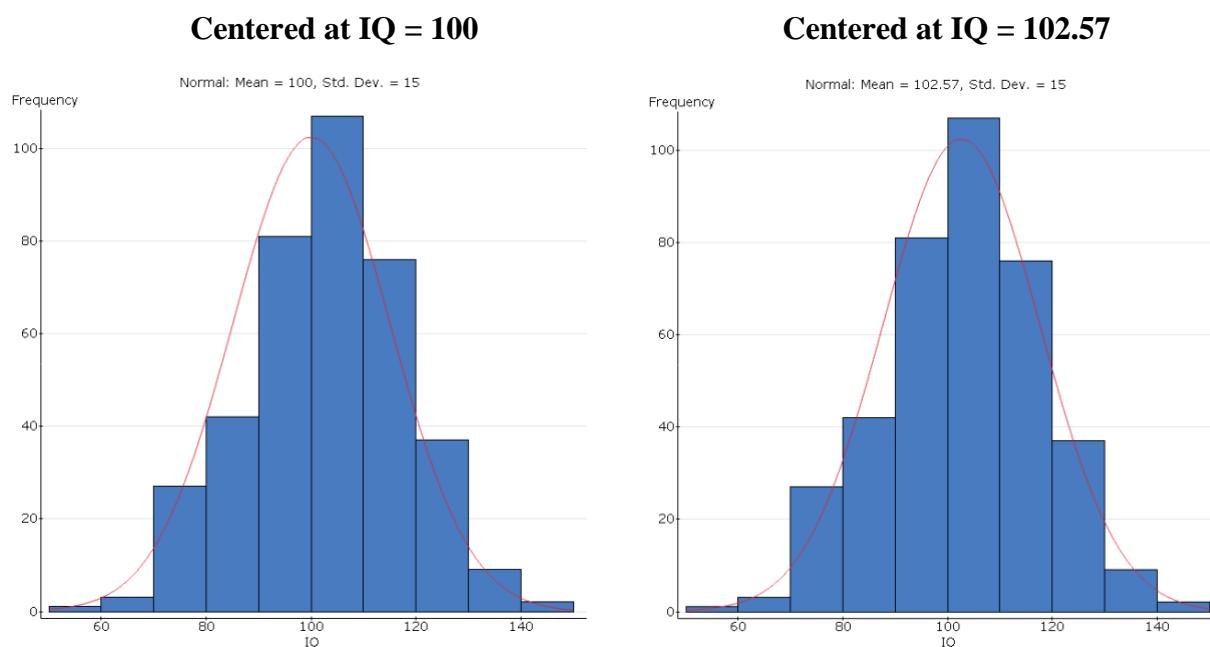
Summary statistics:

Column	n	Mean	Std. dev.	Median	Range	Min	Max	IQR
IQ	385	102.57	15.00	103	88	58	146	19

10th Per.	20th Per.	30th Per.	40th Per.	60th Per.	70th Per.	80th Per.	90th Per.
83	90	95	99.5	107	110	114.5	121



- Researchers at this point might ask the question, “***If Ivy League students really had an average IQ of 100, how unlikely would it be to sample 385 of them and get a mean IQ of _____?***”
- Researchers at this point might try to draw a smooth curve that describes the general distribution of IQ scores at Ivy League schools. The **smooth curve** is the model. It can be represented by a **mathematical equation**, but in this case, that is a bit beyond the scope of Math 127:

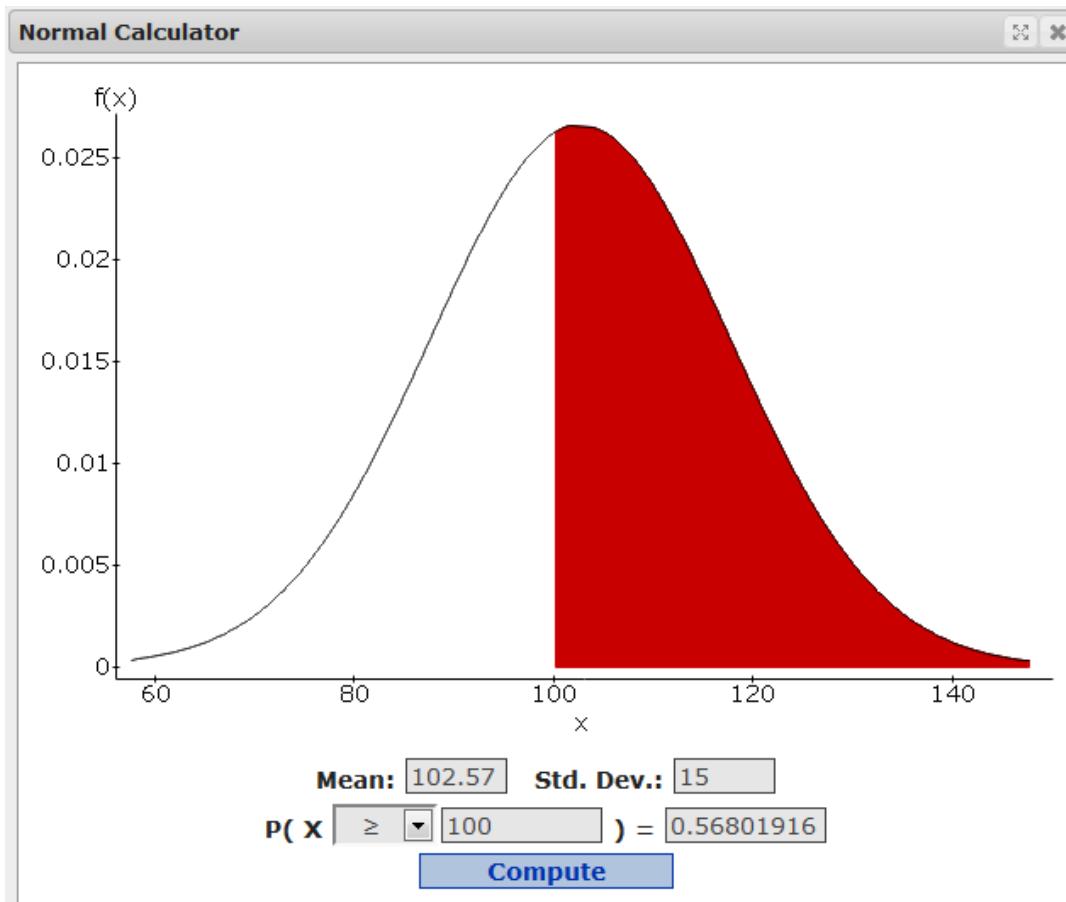


- As you can see, both models are _____. Models are generalizations, so they will be _____. We just hope for our model to be _____.
- A big part of the field of Statistics is choosing the best model. With the data collected by the researchers, the model on the right, centered at IQ = 102.57, seems to be the slightly better choice.
- Realize this:** If the researchers were to collect new data (they won't), the histogram and summary statistics would very likely be _____ from what we originally collected. This would lead to a different model.

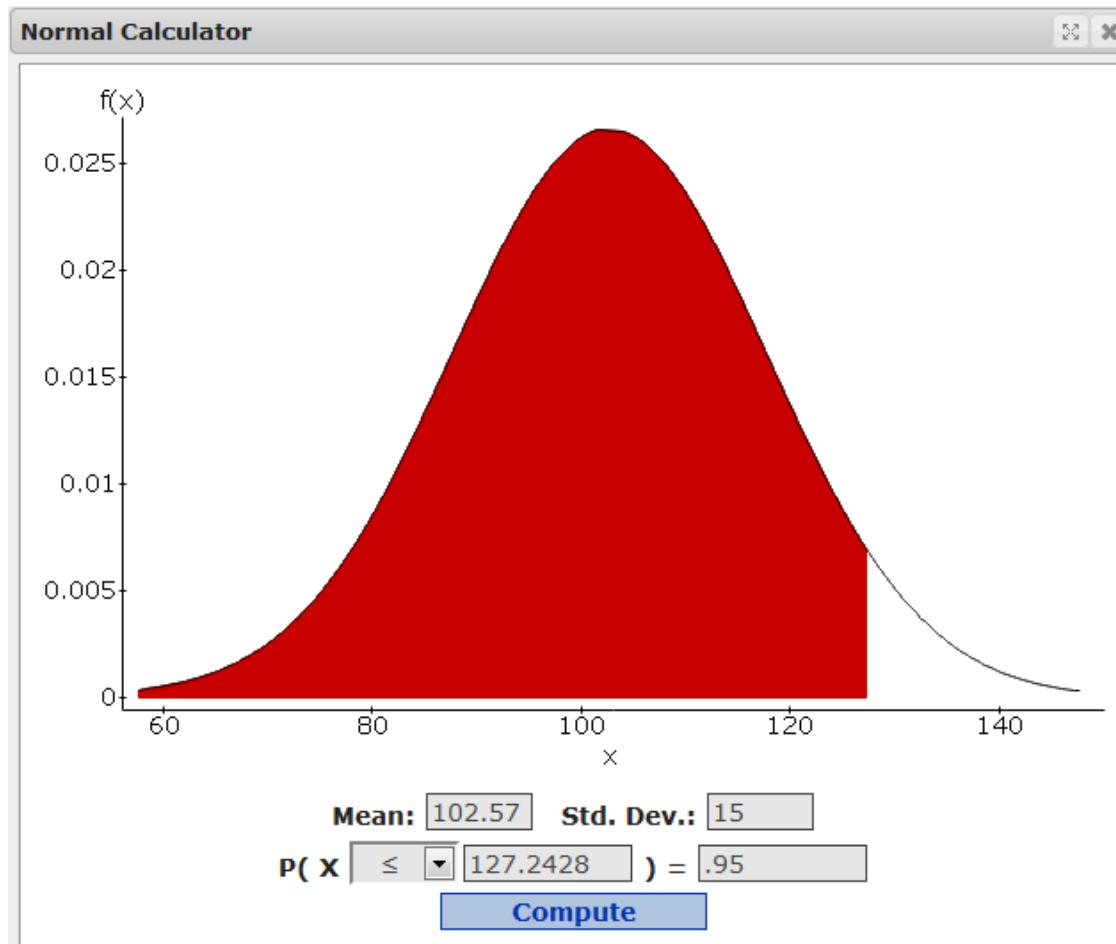
What are Models Good For?

- The answer is _____ . If the model is _____ , our _____ should be _____ .

Question: Presuming the Normal model centered at IQ = 102.57 is a good model, what percentage all of Ivy League students should we expect to have IQs exceeding 100 (the typical average IQ)?



Question: Presuming the Normal model centered at IQ = 102.57 is a good model, what is the IQ score that marks the 95th percentile? In other words, students with an IQ this high or higher are in the top 5%.



Question: Presuming Ivy League students are no smarter than the typical American (in other words, they have a mean IQ of 100 too), what is the probability that researchers, when they draw a random sample of 385 Ivy Leaguers, would actually get a sample mean of 102.57 or one even larger?

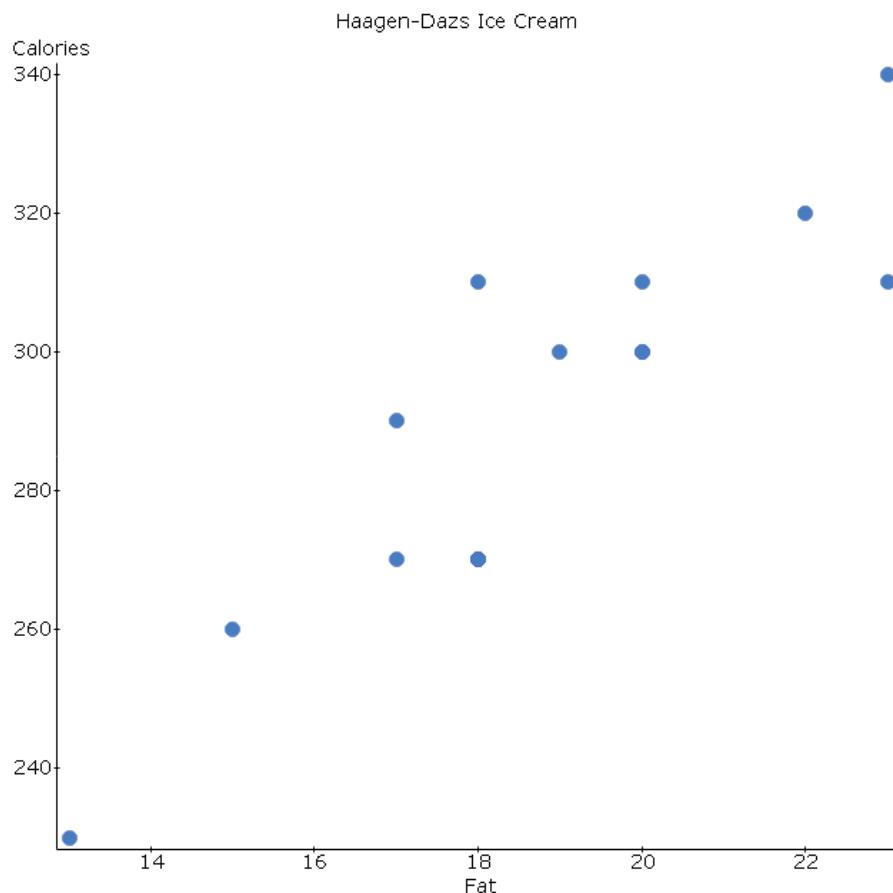
Answer:

Blackboard Video 7-B

Example: Fifteen Haagen-Dazs ice cream varieties were researched on the internet, and “*Fat*” grams were graphed against “*Calories*”. A dietitian was interested in creating a model to predict calorie count per serving based on the fat grams.

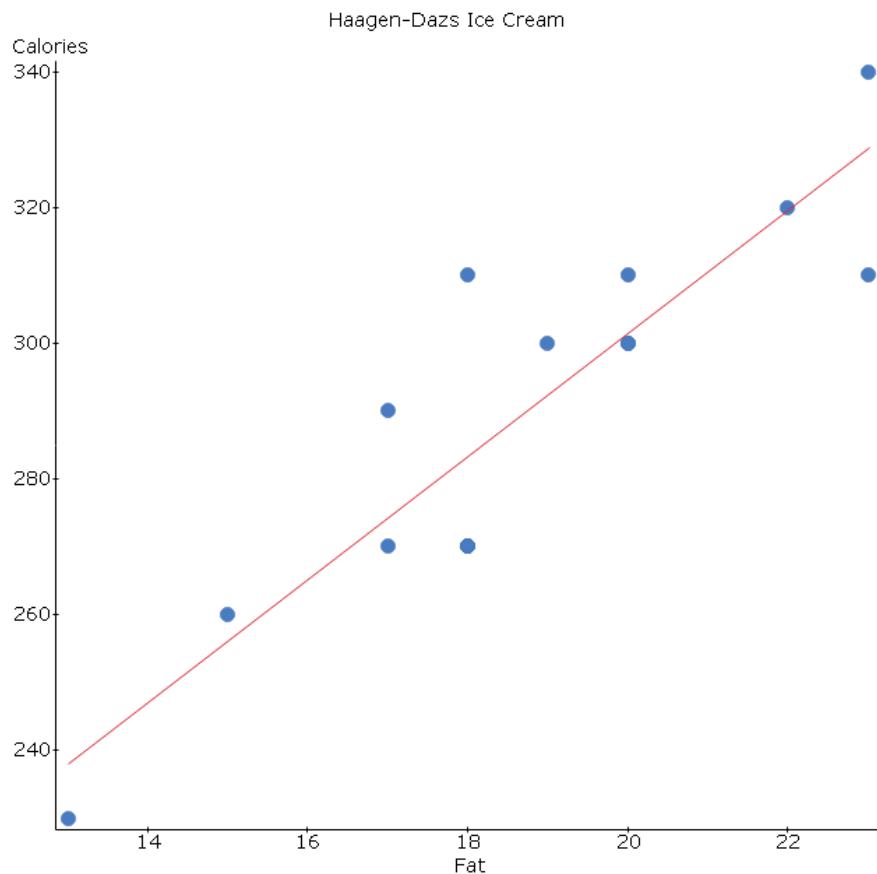
Some of the data is presented in the table below:

Flavor	Vanilla	Butter Pecan	Rum Raisin	Vanilla Swiss	Almond Hazel
Fat Grams	18	23	17	20	22
Calories	270	310	270	300	320



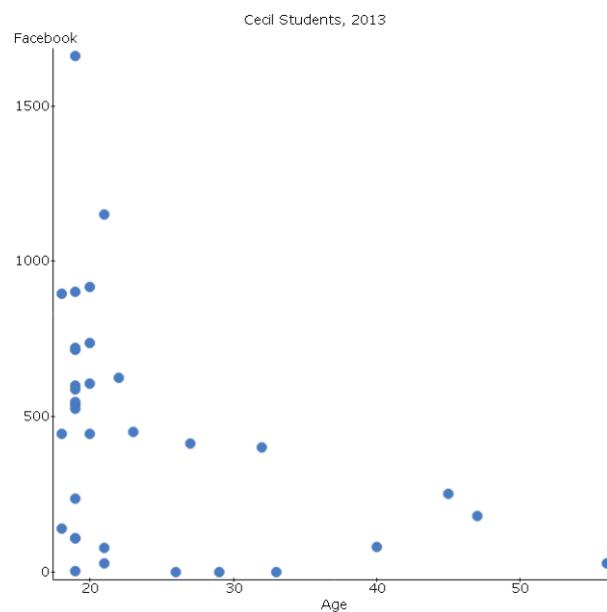
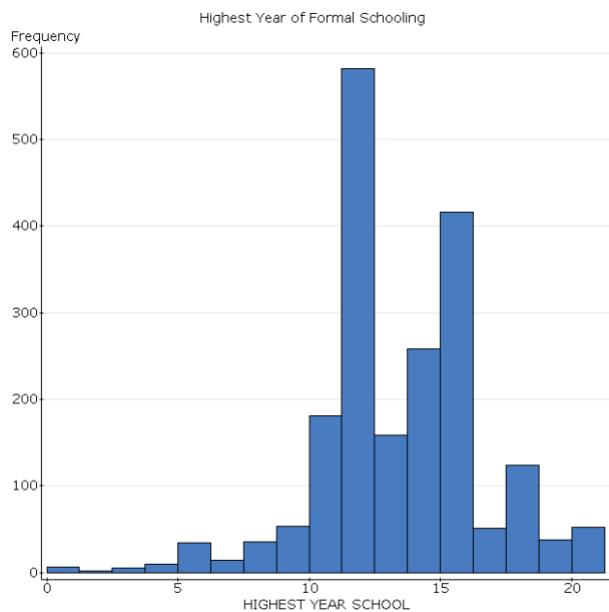
- Because there is a clear relationship, modeling the data with a straight line would be appropriate.
- The model won't be _____, because most every point won't fall exactly on the line, but the model should still prove to be _____ for prediction and for understanding the relationship between the variables.

- Here is the scatterplot with the best linear equation plotted along with the data:



- What might the dietician predict for the calorie count if a new flavor of ice cream has 15 grams of fat?
- Should we use this linear equation to predict the calorie count for a new fat-free ice cream?

- In Math 127, one model we will primarily study is the _____ model for **one quantitative variable**. We must be sure to _____ our data first with a _____ before deciding to use this model – if the data doesn't look _____ and _____, then a Normal model won't apply.
- When looking at the relationship between **two quantitative variables**, the _____ model will often be a good choice. Again, we will first graph our data with a _____ to be sure the relationship is straight and reasonably strong.
- We must remember that models are merely that – generalizations of the real world based on the limited data that we've just happened to collect with an experiment, a survey, a random sample, or by some observational study we found on the internet.
- **Not all data can be successfully modeled**. Below are two examples where it would be quite tough to find a useful and simple model. On the left, we have from the “**General Social Survey 2008**” respondents “**Highest Year of School**”. On the right, we surveyed Cecil students in 2013 and explored the relationship between their “**Age**” and “**Number of Facebook Friends**”.



Lesson 8: Linear Models – Graphs and Summary Statistics

Blackboard Video 8-A

To motivate our discussion of scatterplots, correlation, and linear regression, we will follow an example throughout Lessons 8 and 9.

Example: The U.S. Government, through its Direct Loan program, makes loans directly to college students. We are investigating only subsidized loans – that is, the interest is paid by Uncle Sam until you are out of school.

Data was collected during a recent 4th quarter (Oct – Nov – Dec), and a simple random sample of 62 schools was drawn from the population of all schools granting loans. Cecil College was added to the dataset for local interest.

The variables we will be looking at “**Number of Loans**” and “**Loan Value**” for each of the schools in our dataset. For example, Cecil College issued 178 loans at a total value of \$295,818.

For planning purposes, the U.S. Government wants to explore the relationship between “**Number of Loans**” and “**Loan Value**” – it is easy to estimate how many loans will be granted any given semester, as those trends are fairly stable. The total loan value is more difficult since schools have varying tuition rates, tuition increases, enrollments, and so on. By the end of Lesson 9, we will come up with a model to predict “**Loan Value**” based on “**Number of Loans**”.

Row	School	Number of Loans	Loan Value
1	UNIVERSITY OF TEXAS AT ARLINGTON	12643	2.8496868E7
2	WEST VIRGINIA UNIVERSITY	12468	3.0520856E7
3	STATE UNIVERSITY OF NEW YORK AT BUFFALO	9799	2.7752068E7
4	UNIVERSITY OF WASHINGTON - SEATTLE	9165	1.8446376E7
5	UNIVERSITY OF CALIFORNIA, SANTA BARBARA	6705	1.0984619E7
6	UNIVERSITY OF SOUTH ALABAMA	5846	1.3941994E7
7	CLEMSON UNIVERSITY	5459	1.3782979E7
8	MONTANA STATE UNIVERSITY - BOZEMAN	4681	1.0565853E7
9	UNIVERSITY OF CENTRAL MISSOURI	4583	1.0652497E7
10	UNIVERSITY OF NORTH FLORIDA	4263	1.0331304E7
11	UNIVERSITY OF MASSACHUSETTS - DARTMOUTH	3756	8759167
12	PENNSYLVANIA COLLEGE OF TECHNOLOGY	3603	7250301
13	UNIVERSITY OF TENNESSEE - CHATTANOOGA	3439	8094439
14	RADFORD UNIVERSITY	3433	7849140
15	BROWARD COLLEGE	3295	6132762

The Roles of the Two Variables

When looking at the relationship between **two quantitative variables**, first decide which variable is the “variable of interest” –

We call the variable of interest _____ or the _____

Direct Loans Example: _____

We call the other variable _____ or the _____

Direct Loans Example: _____

Example: Suppose we would like to predict a person’s “**Income**” based on their “**Age**”. Identify the response variable and the explanatory variable.

Example: Cecil students obtained the price of a textbook they had needed for the current semester. Students recorded the price for their book at the “**Bookstore**”, on “**Amazon.com**”, and on “**EBay**”. The goal of the investigation was to determine which website most accurately predicted the “**Bookstore**” price. Identify the response variable and the explanatory variables.

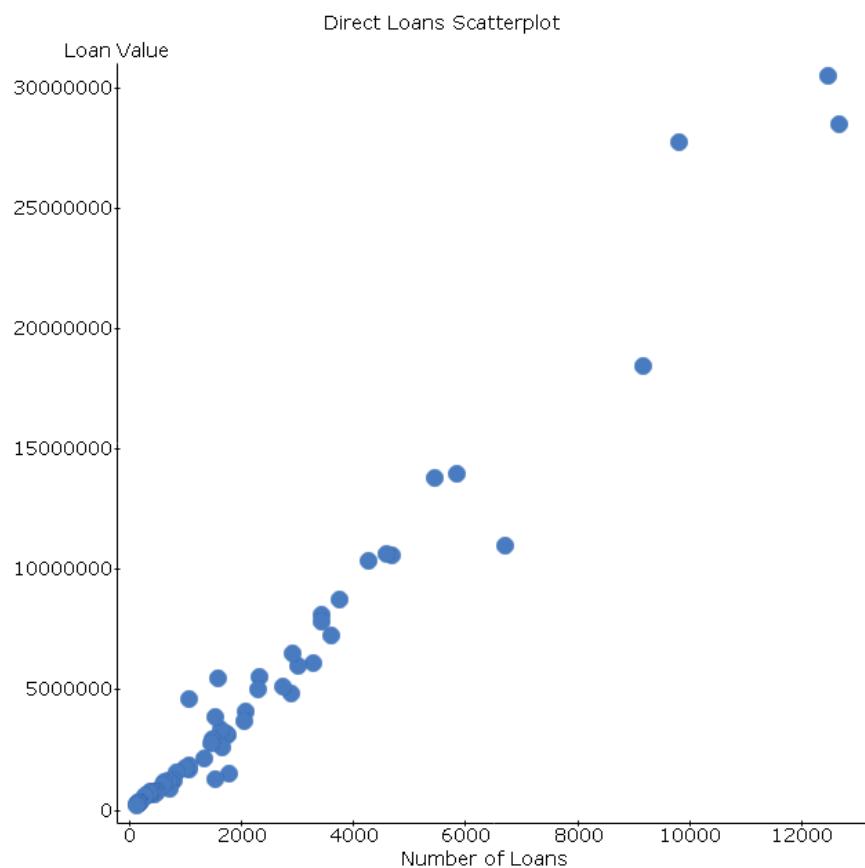
- Whatever we are predicting is the _____ variable.
- We use the explanatory variable, x , to _____ the values of the response variable, y .
- We **must** properly name the variables and understand their roles in a regression analysis. If you switch the variables, the whole analysis will be backwards and faulty.

Blackboard Video 8-B

Graphing Two Quantitative Variables

Always make a graph; here the right one is a _____.
The explanatory variable goes on the x -axis, and the response variable on the y -axis.

Example: For “**Direct Loans**”, “**Number of Loans**” is the explanatory variable, so it goes on the _____ axis. “**Loan Value**” is the response variable, so it goes on the _____ axis. Comment on the scatterplot.

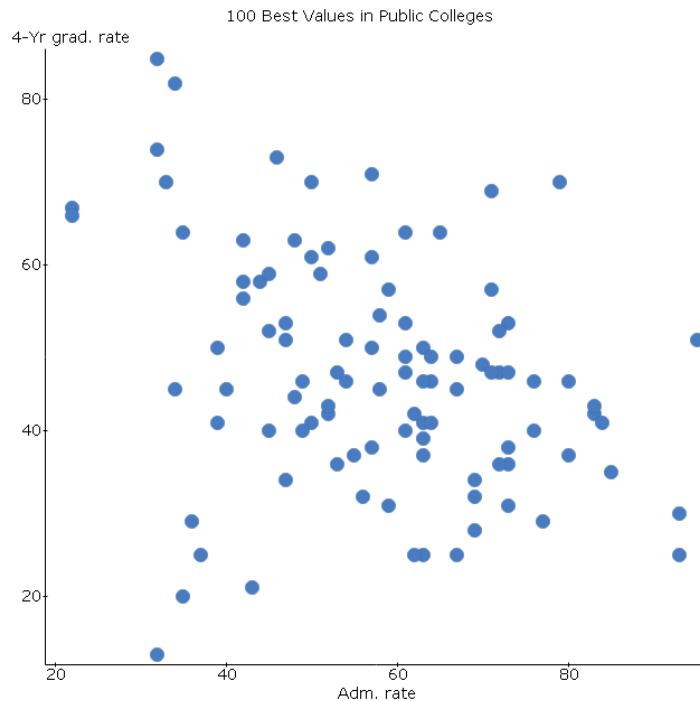


Describing the Relationship Between Two Quantitative Variables

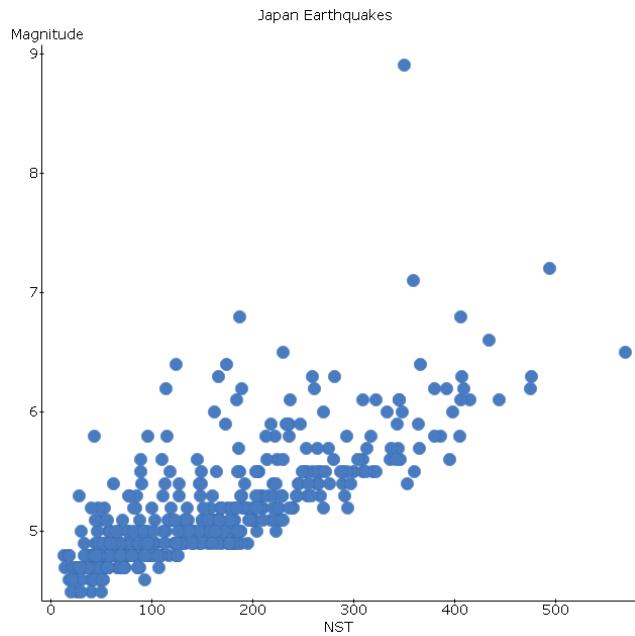
- When asked to describe the relationship between two quantitative variables, you must mention 4 things every time:
- Form** can be _____, _____, or _____.
 - Direction** can be _____ or _____.
 - Strength** is a measure of how much scatter. Though highly context specific, we will use words like _____, _____, or _____.
 - Look for the **unusual features**. Are there clusters of points that require further investigation or are there a few points that could be considered outliers? If so, note them down and proceed with caution. Are there no unusual features? Then say that.

Example: For each scatterplot, describe the relationship between the two variables by noting down the 4 points mentioned above.

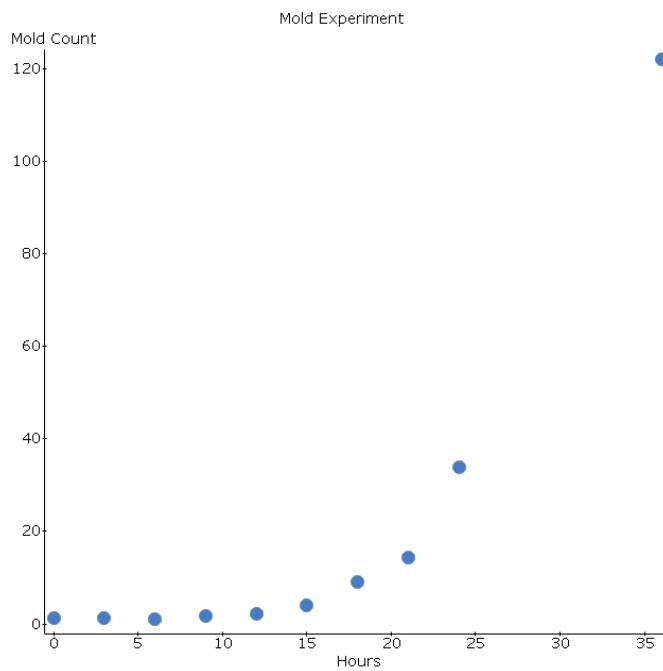
- From the “**Kiplinger’s 100 Best Values in Public Colleges**” dataset, we are looking at using the “**Admission Rate**” to predict the “**4-Year Graduation Rate**”.



- b. From the “**Honshu Japan Earthquake**” dataset, we are looking at using the “**Number of Stations**” reporting earthquake activity to predict the “**Magnitude**” of an earthquake.



- c. From the “**Mold Count**” dataset, we are looking at using the “**Hours**” since beginning to grow the mold to predict the “**Mold Count**” in a Petri dish.



Blackboard Video 8-C

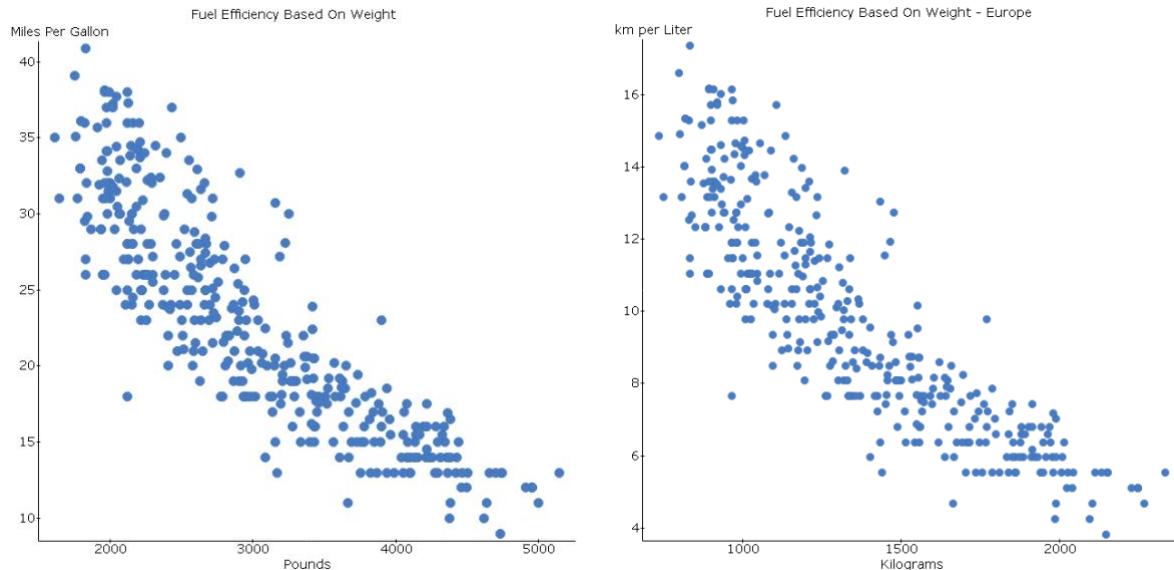
Calculating the Correlation Coefficient

If the association is linear, then the next step is to quantify the strength and direction of the linear relationship. The statistic that does this is called the correlation coefficient, _____.

- To understand the **correlation coefficient**, we first must understand that for our measure of strength and direction, units won't matter.

Example: Data collected on 370 automobiles included $x = \text{"Weight"}$ measured in pounds and $y = \text{"Miles Per Gallon"}$.

On the left are American measurements. On the right, European measurements. Same cars, just transported across the ocean. The correlation must be the same, even though we changed our units of measurement.



- In both cases, the correlation coefficient is _____. Changing the units didn't change the value of the correlation.
- Another way we can change the units of measurement is to take our data values and convert them to _____. This is the way correlation is computed.

Example: For “**Direct Loans**”, the first step to calculating the correlation is to convert our data values into the z -scores.

Recall: How do you calculate a z -score?

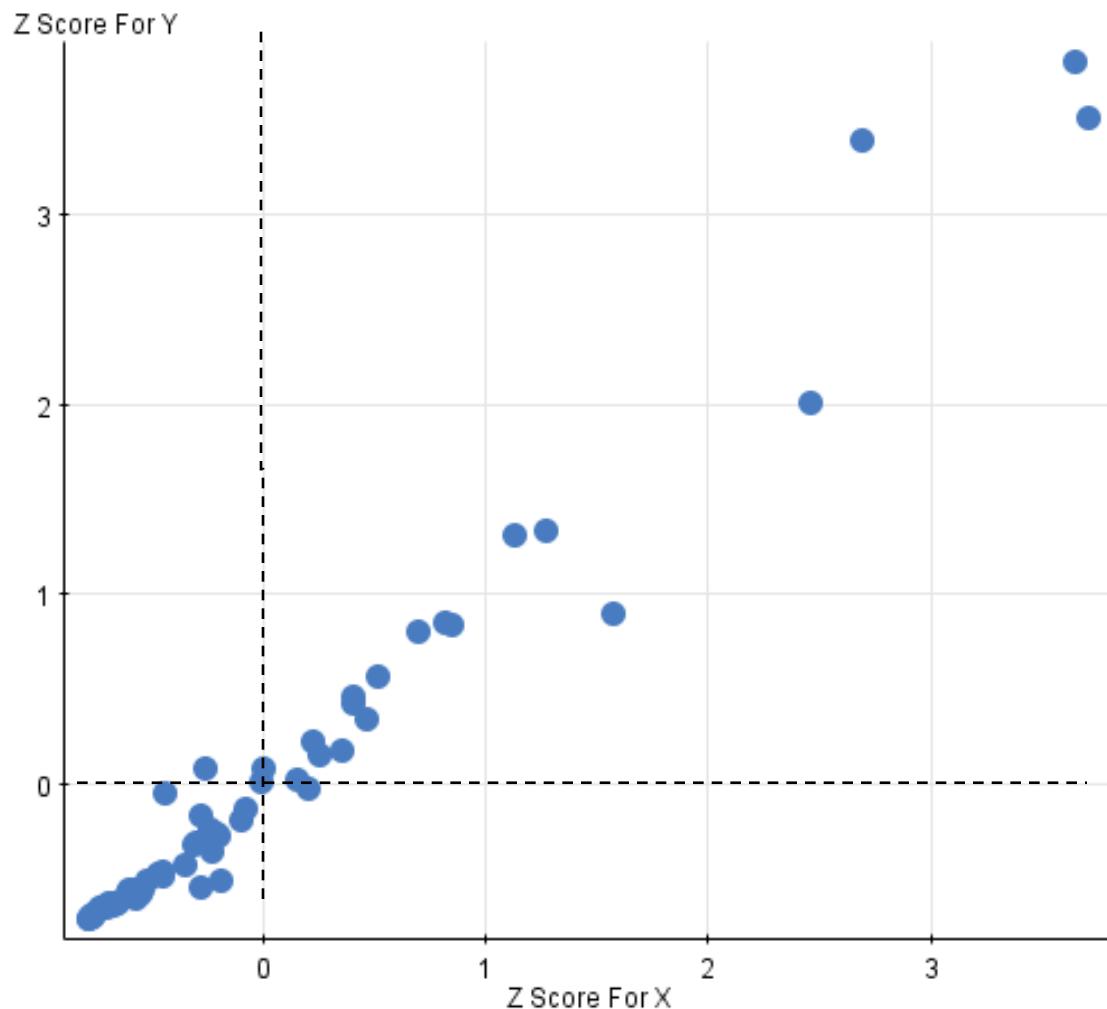
Summary statistics for “**Direct Loans**”:

Column	n	Mean	Standard Deviation
Number of Loans	63	2,308.56	2,787.52
Loan Value	63	\$4,987,964	\$6,704,153

“**Direct Loans**” data with z -scores computed:

StatCrunch	Edit	Data	Stat	Graphics	Help
Row	School		Number of Loans	Loan Value	Z-Score for X
1	UNIVERSITY OF TEXAS AT ARLINGTON		12643	2.8496868E7	3.7073958
2	WEST VIRGINIA UNIVERSITY		12468	3.0520856E7	3.644616
3	STATE UNIVERSITY OF NEW YORK AT BUFFALO		9799	2.7752068E7	2.687134
4	UNIVERSITY OF WASHINGTON - SEATTLE		9165	1.8446376E7	2.4596918
5	UNIVERSITY OF CALIFORNIA, SANTA BARBARA		6705	1.0984619E7	1.577187
6	UNIVERSITY OF SOUTH ALABAMA		5846	1.3941994E7	1.2690277
7	CLEMSON UNIVERSITY		5459	1.3782979E7	1.1301945
8	MONTANA STATE UNIVERSITY - BOZEMAN		4681	1.0565853E7	0.8510935

- Below, we have created a scatterplot for “**Direct Loans**”, but instead of graphing “**Number of Loans**” versus “**Loan Value**”, we have graphed the z -scores instead.
- Develop the ideas to calculate the correlation:



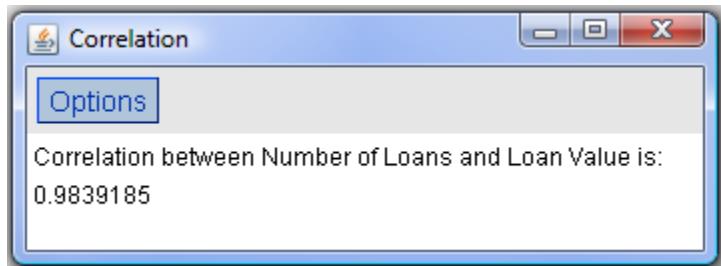
- Now, for **Direct Loans**, we clearly have a positive association.
- With a very strong and positive relationship, most of the points are in the _____ and _____ quadrants.
- Take a point in the **first** quadrant and multiply the z -scores together. The result is:
_____.
- Take a point in the **third** quadrant and multiply the z -scores together. The result is:
_____.
- We have a few points in the **second** and **fourth** quadrants – those products of z -scores turn out to be _____.
- For points further from the origin, the product of z -scores will get _____.
- For points close to the origin, the product of z -scores will get _____.

Net Result: We want a single number that quantifies the strength and direction of the linear association. The correlation coefficient adds up all the z -score products (one for each data point) and then “averages them out”. This number does the job – hopefully you can see why!

Formula: The correlation coefficient, to measure the strength and direction of the relationship between two quantitative variables, is given by:

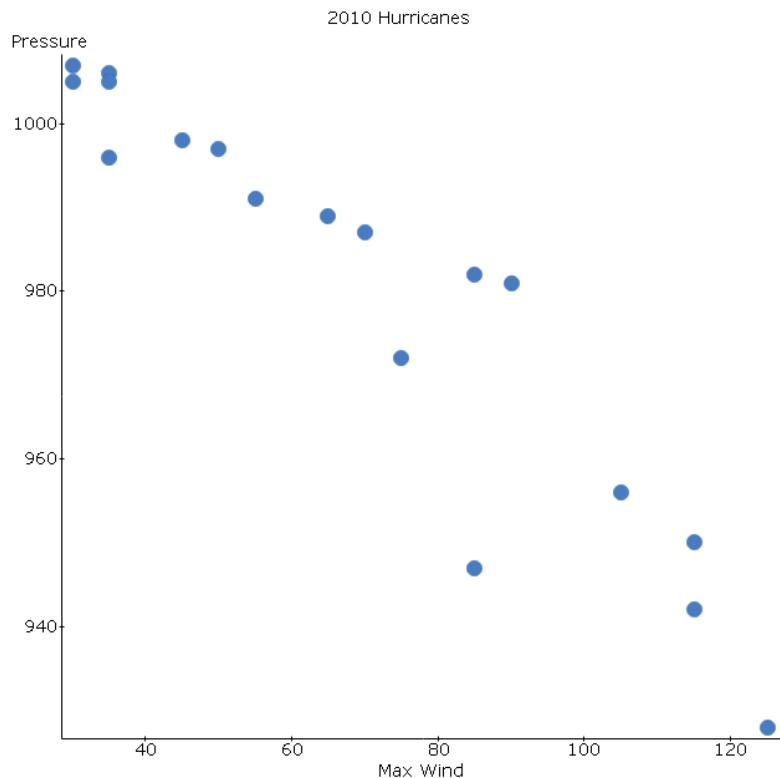
Note: Calculating the correlation by hand is going to be one of the most tedious tasks in Math 127. But don’t worry, we let StatCrunch handle the details.

Direct Loans: StatCrunch tells us that for “*Number of Loans*” and “*Loan Value*”:



Interpret this value:

Example: For the “**2010 Hurricane**” dataset, we are looking at the relationship between $x =$ “**Max Wind**” and $y =$ “**Pressure**”. Discuss the idea behind calculating the correlation coefficient of $r = -0.937$.



Blackboard Video 8-D

Facts About Correlation

- **Correlation** is a number that describes the _____ and _____ of a linear relationship between two _____ variables.
- **Association** is a deliberately vague term describing the relationship between two variables. It only denotes that some sort of relationship exists.
- **Causation** is a word that indicates a _____ relationship. Scatterplots and correlations do **not** prove causation! Causation is stronger than correlation.
- Correlation only applies to _____ variables! Sure, you can calculate a correlation for Baltimore Raven "**Salary**" and "**Jersey Number**", but the result will be meaningless.
- Correlation is meaningful only if the relationship is _____. Check the scatterplot every time! Just because the computer will give you $r = 0.827$ does not mean that the number is meaningful!
- The correlation coefficient can be radically distorted by the presence of outliers. Check the scatterplot first! Investigate the reasons for the outlier and take the appropriate steps.
- The _____ of r always tells us the _____ of the association.
- The minimum value is _____ and the maximum value is _____.
 - In these cases, the points fall exactly on a _____.
- Correlation has _____. That's why we can transform our data to z -scores before calculating the correlation. Z -scores have no units.

- Correlation is reported as a _____, not a percentage.
- There is no magic number for a “good” correlation – it is context specific.

Words to Describe Certain Correlations

- So that everyone in Math 127 is on the same page, let's use the following words to describe certain correlations:

Blackboard Video 8-E

Hans Rosling Video: “Meaningless and Meaningful Correlations”

1. Famous quote: “Correlation does not imply _____”.
 2. Is the quote true?
 3. What four other factors did the tobacco companies suggest could be the cause of lung cancer?
 4. By eliminating the other possibilities, the researchers were able to eliminate all other factors and make the jump to “cause and effect”. What two causal relationships were established about smoking cigarettes?

Blackboard Video 8-F

Example: For each scenario, think what the scatterplot would look like – direction, form, strength, and unusual feature. Which variable is the response, which is the explanatory? Draw a possible scatterplot.

a. A sample of drivers – blood alcohol content and reaction time.

b. A sample of Cecil College students – time spent on school work and GPA.

Lurking Variables

- Be careful! Sometimes a hidden variable stands behind a relationship and affects both the response and explanatory variable simultaneously.

Example 1: Children with larger feet can read at a higher grade level. Amazing!

Possible Lurking Variable: _____

Example 2: Air conditioner sales have been found to be associated with bank robberies. Also, ice cream sales and murders both move in the same direction.

Possible Lurking Variable: _____

Example 3: Damages attributed by house fires are highly associated with the number of firefighters on the scene.

Possible Lurking Variable: _____

Example 4: A researcher studying violent behavior in elementary-school children asks the children's parents how much time each child spends playing video games and then the teachers rate each child on the level of aggressiveness displayed when playing with other children. The researchers find a moderately strong positive correlation. Name a few possible lurking variables.

Lesson 9: Linear Models – Least Squares Regression

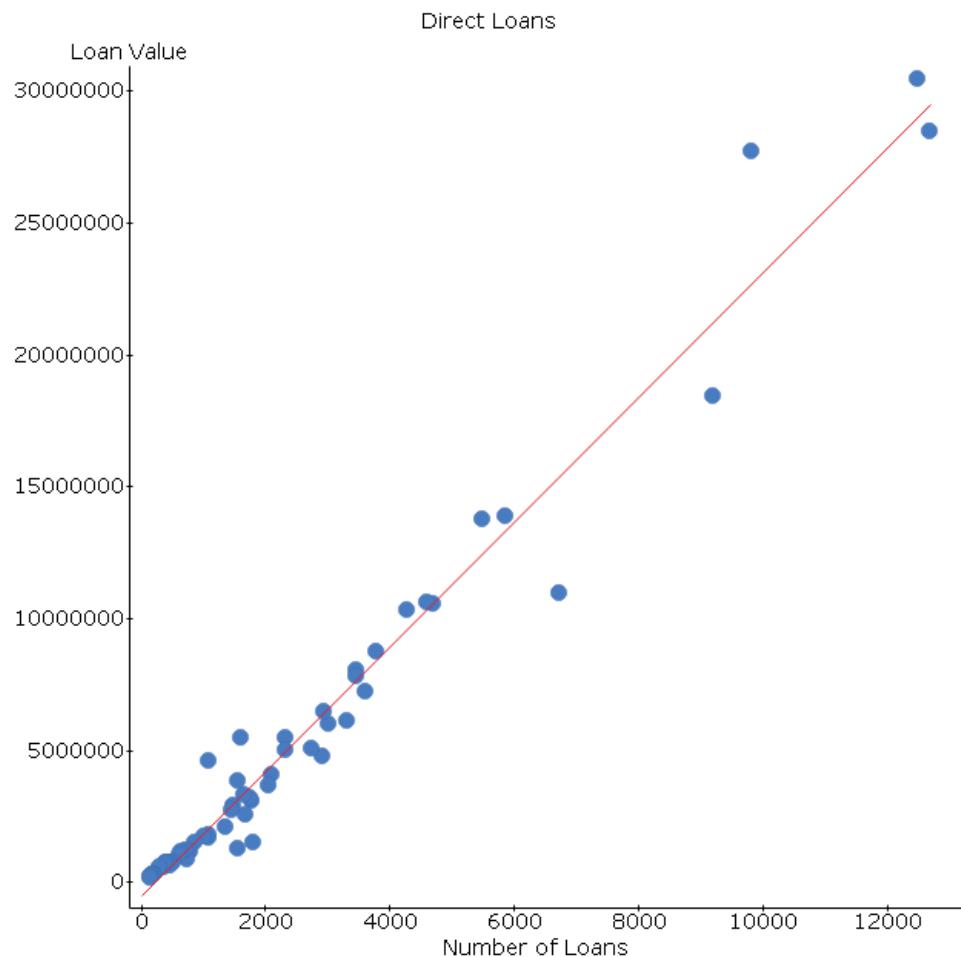
Blackboard Video 9-A

Direct Loans: Previously, we looked at the relationship between “*Number of Loans*” and “*Loan Value*”. Since the relationship was linear, and our $r = 0.984$ indicated a very strong correlation, the next step is to estimate the linear **model** with the **data** we have collected. We hope the model proves useful to the U.S. Government for planning and prediction.

Note: The linear regression equation we create will not match reality _____, but it should still prove to be _____.

Main Idea 1: The Linear Equation is Built to Minimize the Residuals

- Below is our scatterplot, this time with the linear equation included. Draw in a residual:



- The y -value of each data point is called _____.
- The y -value of the line that matches up vertically with each data point is called _____.
- The difference between these two values is the residual.
Residual = _____
- The better the linear equation, the _____ the residuals will be.
- To pick the best line, we use the principle of _____. We pick the line that minimizes the sum of all the residuals (one for each data point).
- There is a fixable problem, though: The points above the line have _____ residuals, while the points below the line have _____ residuals. If we add them all up, they will always cancel each other out and add to _____.
- **Solution:** _____ the residuals before finding the line that minimizes them. Squared numbers are always _____, so least squares linear regression is the process of picking the line that minimizes the squared residuals.
- Another word for residual is _____.
- Written as an expression, we get _____.
- **Note:** Software such as StatCrunch will automatically calculate for us the best line using these principles. We will not have to manually compute every residual for every data point to see if we have discovered the best fitting line.

Blackboard Video 9-B

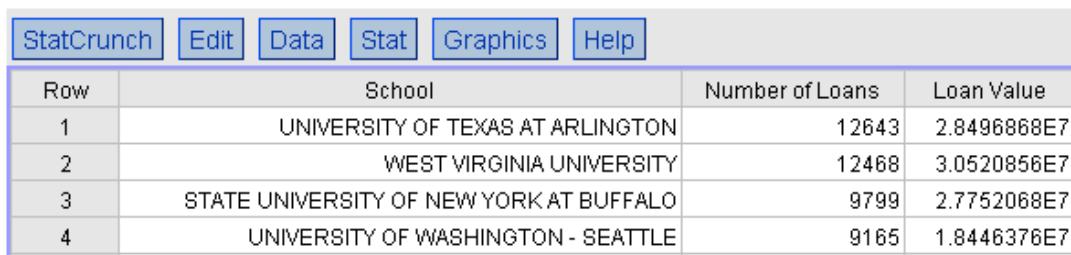
Main Idea 2: Understanding the Equation for the Linear Model

- From **algebra**, the equation of a straight line is:

- Remember that we have **real data**, though, so all our points won't fall exactly on the line.
Our _____ will fall exactly on the line.

- In Statistics, we have different symbols to write the linear regression equation:

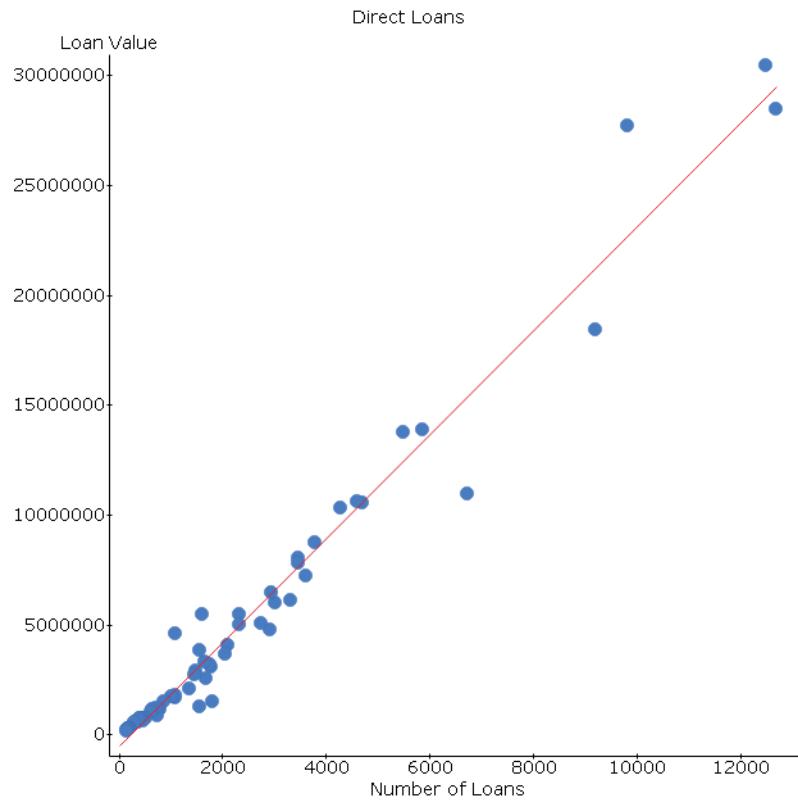
- For the **Direct Loans** example, StatCrunch (using least squares) has given us:



Row	School	Number of Loans	Loan Value
1	UNIVERSITY OF TEXAS AT ARLINGTON	12643	2.8496868E7
2	WEST VIRGINIA UNIVERSITY	12468	3.0520856E7
3	STATE UNIVERSITY OF NEW YORK AT BUFFALO	9799	2.7752068E7
4	UNIVERSITY OF WASHINGTON - SEATTLE	9165	1.8446376E7

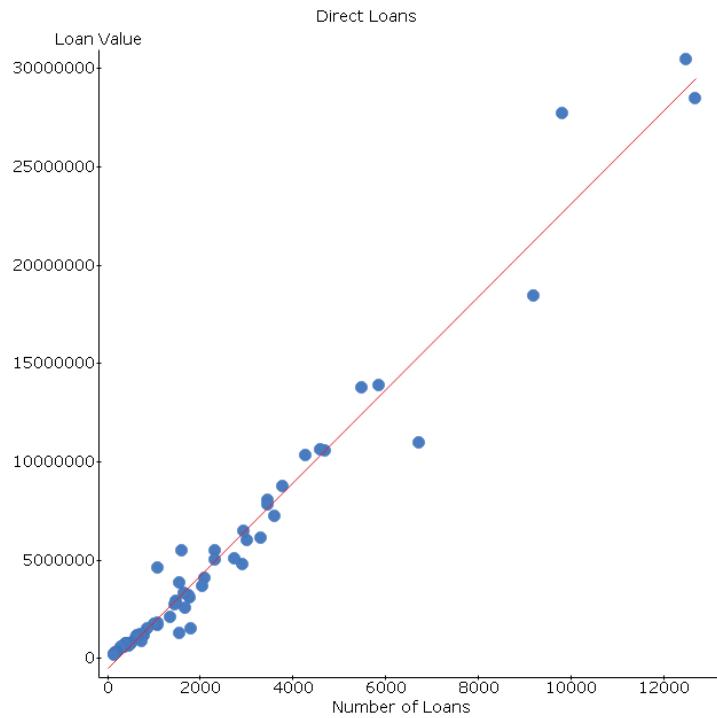
$$\text{Loan Value} = \$2366(\text{Number of Loans}) - \$474,958$$

- For the **Direct Loans** example, the slope of the line is _____. Interpret this value with a sentence in the context of the problem:



- In general, we interpret the value of the slope like this:
- In general, the slope always expressed in _____.
 - When asked to interpret the slope, you always must write a sentence in the context of the problem. For **Direct Loans**, we must speak about the change in the "**Loan Value**" as we increase the "**Number of Loans**" by one.

- For the **Direct Loans** example, the value of the y -intercept is: _____
- Algebraically**, this is the value the line takes when $x = 0$. For **Direct Loans**, the point $x = \text{“Number of Loans”} = 0$ and $y = \text{“Loan Value”} = -\$474,958$ falls on our linear model.



- Every time we create a linear equation, there will be the y -intercept to deal with. Whether or not the y -intercept has meaning in the context of the problem depends on a few things:

- Do we have any data at $x = 0$ or **very** close to $x = 0$?

If we don't, then the y -intercept falls outside the scope of our data and interpreting its value would be extrapolating, which is a bad thing.

For **Direct Loans**, we do not have any schools with $x = 0$ Loans. The smallest value in the dataset is the Francis Tuttle Technology Center with $x = 128$ Loans.

- Does the value of the y -intercept make sense in the context of the problem?

Our model is predicting "**Loan Value**" to be $-\$474,958$ if we were to have a school grant 0 loans. This value makes no sense in the context of the problem.

- Therefore, in our **Direct Loans** example, the point on the line of $x = 0$ and $y = -\$474,958$ **has no meaning** other than it's the point where the line crosses the y -axis.

Blackboard Video 9-C

Example: For each scenario, interpret the slope and y -intercept based on the context of the problem.

- a. A college instructor created a linear equation to predict $y = \text{“Student’s Final Grade”}$ based on $x = \text{“Number of Absences”}$. From the 58 students, the equation was:

$$\text{Student's Final Grade} = 96.77 - 5.19(\text{Number of Absences})$$

Slope:

y -intercept:

- b. A geologist collected data from 77 elevations up in Rocky Mountains in Montana during last December. Elevation (1000s of feet above sea level) was used to predict temperature (degrees Fahrenheit). Here is the linear equation:

$$\text{Temperature} = 60 - 5.6(\text{Elevation})$$

Slope:

y -intercept:

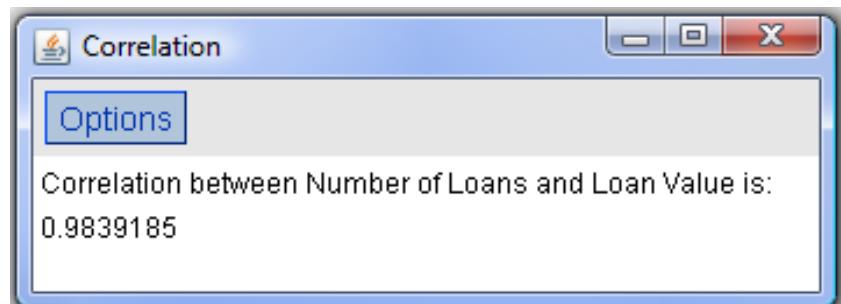
Blackboard Video 9-D

Main Idea 3: Formulas for the Slope and the y-Intercept

- For Direct Loans, here is a snapshot of the data, the linear equation, and the summary statistics:

		StatCrunch	Edit	Data	Stat	Graphics	Help
Row	School			Number of Loans	Loan Value		
1	UNIVERSITY OF TEXAS AT ARLINGTON			12643	2.8496868E7		
2	WEST VIRGINIA UNIVERSITY			12468	3.0520856E7		
3	STATE UNIVERSITY OF NEW YORK AT BUFFALO			9799	2.7752068E7		
4	UNIVERSITY OF WASHINGTON - SEATTLE			9165	1.8446376E7		

Column	<i>n</i>	Mean	Standard Deviation
Number of Loans	63	2,308.56	2,787.52
Loan Value	63	\$4,987,964	\$6,704,153



$$\text{Loan Value} = \$2366(\text{Number of Loans}) - \$474,958$$

- There are formulas to compute the slope and the *y*-intercept, though many times we won't have to use them if we have the data loaded into software.
- The formulas are interesting in their own right because they are based on three of the most important summary statistics we know: the _____, the _____, and the _____.

Formula for the Slope:**Calculate the Slope for Direct Loans:**

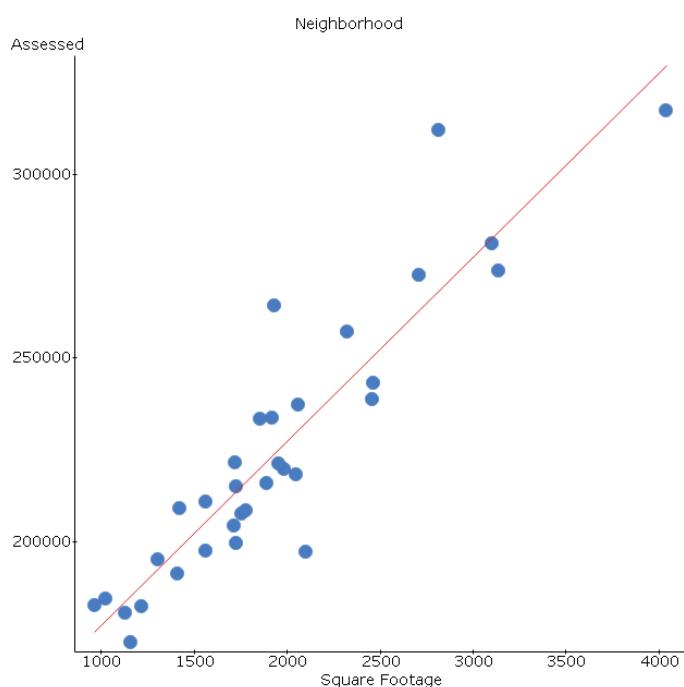
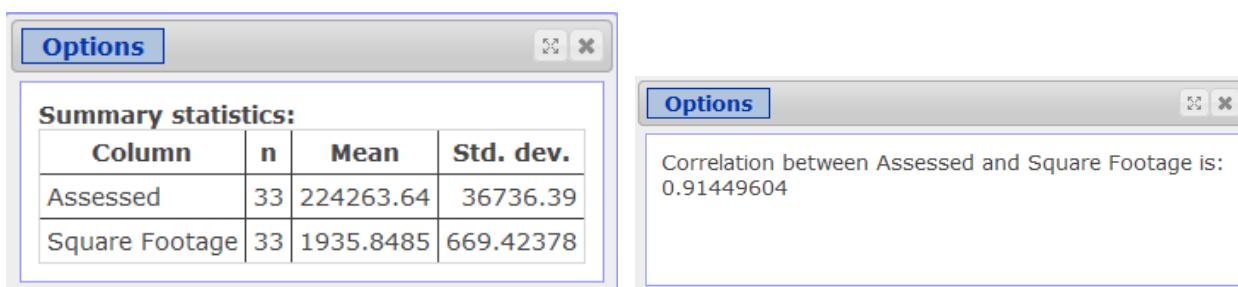
- The formula for the y -intercept is based on this fact: On every linear regression equation, we are guaranteed to find the point (\bar{x}, \bar{y}) .

Formula for the y -Intercept:**Calculate the y -Intercept for Direct Loans:**

Blackboard Video 9-E

Example: In the “Neighborhood” dataset, we would like to create a linear equation to predict the “Assessed” value of a house based on its “*Square Footage*”.

Row	Address	Assessed	Zillow Value	Square Footage	Lot
1	202 Friendship	233700	212200	1918	11761
2	204 Friendship	218400	211500	2048	27094
3	206 Friendship	243100	223300	2460	11935
4	208 Friendship	207700	201900	1752	10802
5	210 Friendship	219900	199100	1984	11325



a. Determine the linear regression equation based on the summary statistics from the previous page.

b. Interpret the value of the slope with a sentence in context.

c. Interpret the value of the y -intercept with a sentence in context.

Blackboard Video 9-F

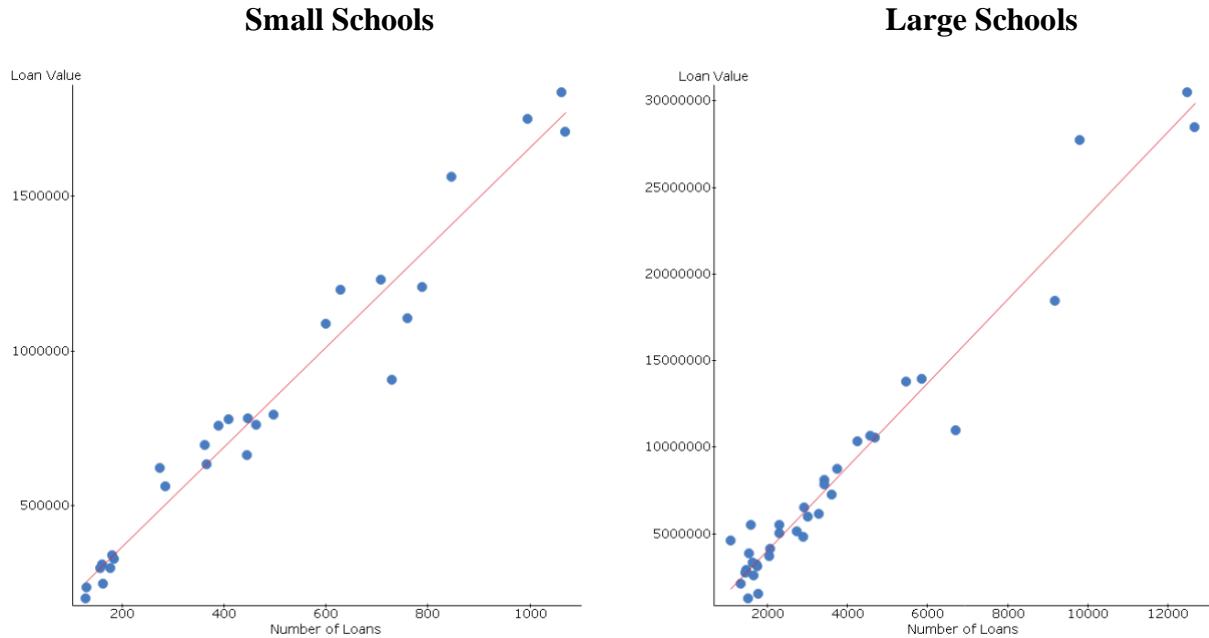
Main Idea 4: Using The Linear Equation to Make Predictions

- The main reason we create a linear equation to summarize the relationship between two variables is to make predictions. Recall that for **Direct Loans**, we have:

$$\text{Loan Value} = \$2366(\text{Number of Loans}) - \$474,958$$

- Suppose that last quarter at the University of Delaware, 4222 loans were issued. What would our equation predict the total “**Loan Value**” to be?
- How much in “**Loan Value**” might we expect Cecil College to grant during the next quarter if they were to grant 178 loans?
- Looking at our **data**, we find that in a recent quarter, Cecil College granted 178 loans totaling \$295,818. This is a far cry from what our equation actually predicted.
- Our **Direct Loans** equation has some problems for the smaller schools because our dataset has huge schools like U. Texas and W. Virginia University lumped together with tiny community colleges like Cecil College.
- To improve our analysis and have more realistic predictions, one suggestion would be to run separate linear regressions for the large universities and the small community colleges. We have made the mistake of lumping all our data together when inherently it is coming from two very different groups.

- To further the investigation, we split the colleges into two groups, “***Small Schools***” and “***Large Schools***”:



$$\text{Loan Value} = \$1616(\text{Number of Loans}) + \$43,798 \quad \text{Loan Value} = \$2419(\text{Number of Loans}) - \$802,634$$

- Now use the small school linear regression equation to predict the “***Loan Value***” for Cecil College with $x = 178$ loans.

Extrapolation

- Extrapolation is using your equation to predict outside the x -values that you actually have data for. Extrapolation is risky and in general, not a good idea.
- For our “***Small Schools***”, it is OK to predict for any “***Number of Loans***” from about $x = 100$ to about $x = 1100$ or 1200. We wouldn’t predict for any x under 100 or above 1200.
- For our “***Large Schools***”, it is OK to predict for any “***Number of Loans***” from about $x = 1200$ to about $x = 13,000$. Outside those bounds is **extrapolating**.

Blackboard Video 9-G

Example: Summary statistics for the data relating the latitude and average January temperature for 50 large U.S. cities are given below. The two variables look to be linearly associated, based on the scatterplot, and the correlation coefficient was $r = -0.894$.

Variable	Mean	Standard Deviation
Latitude	38.02	6.42
January Temperature	27.55	15.49

- d. Explain in context what the y -intercept means.
- e. Elkton sits at 39 degrees north. Predict the mean “*January Temperature*” there.
- f. If a city has a positive residual, what does that mean?
- g. What “*Latitude*” values can we use our model for prediction?

Blackboard Video 9-H

Main Idea 5: Evaluating the Quality of Your Linear Regression Equation

- When fitting a linear equation to a dataset with an explanatory variable, x , and a response variable, y , there are a few conditions on the data that we must check before completing our analysis.

1. The relationship must be _____.

Why is this important?

2. When looking at a scatterplot with the linear equation included, the data points must have _____ about the line.

Why is this important?

3. When looking at a scatterplot, there should not be any data points that are extremely separated from the main cluster of points.

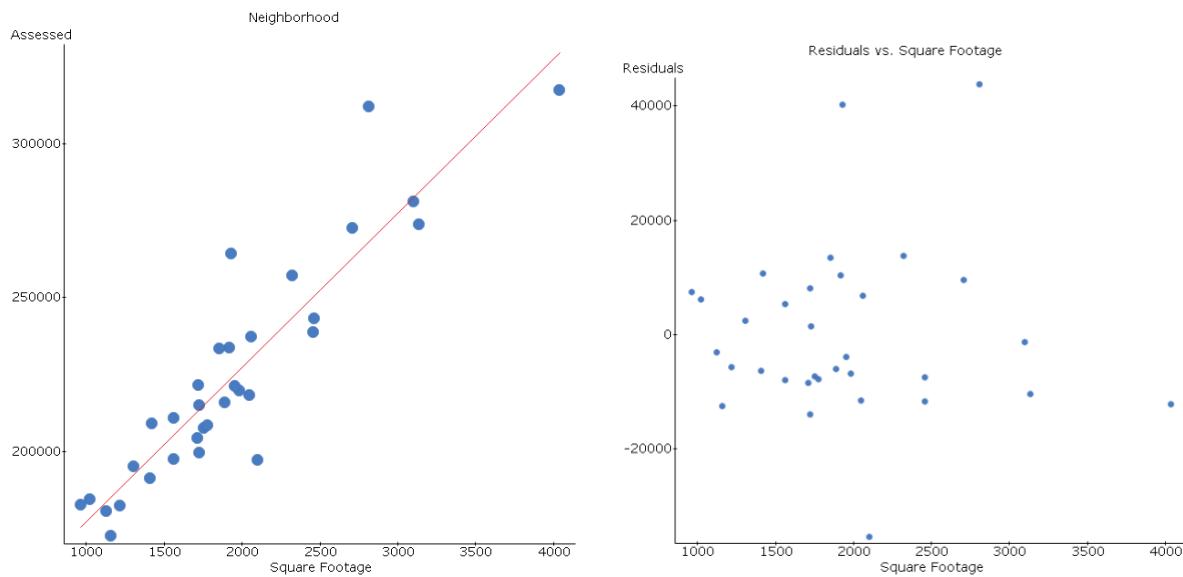
Why is this important?

- To check that our data meets these requirements, we must carefully analyze **two** scatterplots.
- The **first** scatterplot is the one we are familiar with: explanatory variable on the x -axis and response variable on the y -axis. We look for three things:
 - The **second** scatterplot is called a _____ plot: explanatory variable on the x -axis but now we put the **residuals on the y-axis**. We look for three things here too:
 1. In a residual plot, since we are graphing the x -variable versus the residuals from a linear equation, we hope to see a _____ of points.
Remember that the residuals (or errors) are what is left over **after** the linear equation has been accounted for. If our choice of linear was the correct choice, **the residual plot should look random.**
 2. In a residual plot, we hope to see _____ in the data points as we move left to right across the graph.
Remember that the residuals are required to have the same variation about the line for all of the data we collected. It is easier to see on a residual plot than the original scatterplot.
 3. In a residual plot, we can locate any very large or very small residuals by looking for points near the top and bottom of the graph. These data points need to be carefully investigated.
Remember, large residuals (positive or negative) are points where the actual y -value differs substantially from what the regression equation would have predicted.

Blackboard Video 9-I

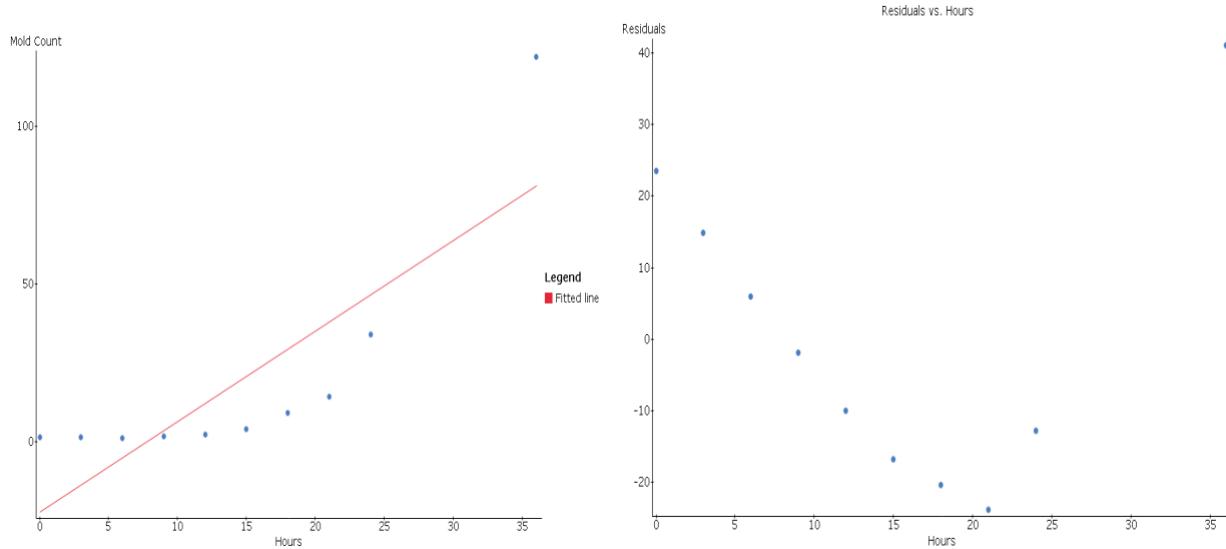
Example: Let's return to the “**Neighborhood**” dataset. We would like to create a linear equation to predict the “**Assessed**” value of a house based on its “**Square Footage**”.

Comment below on the conditions needed to proceed with a linear regression analysis.

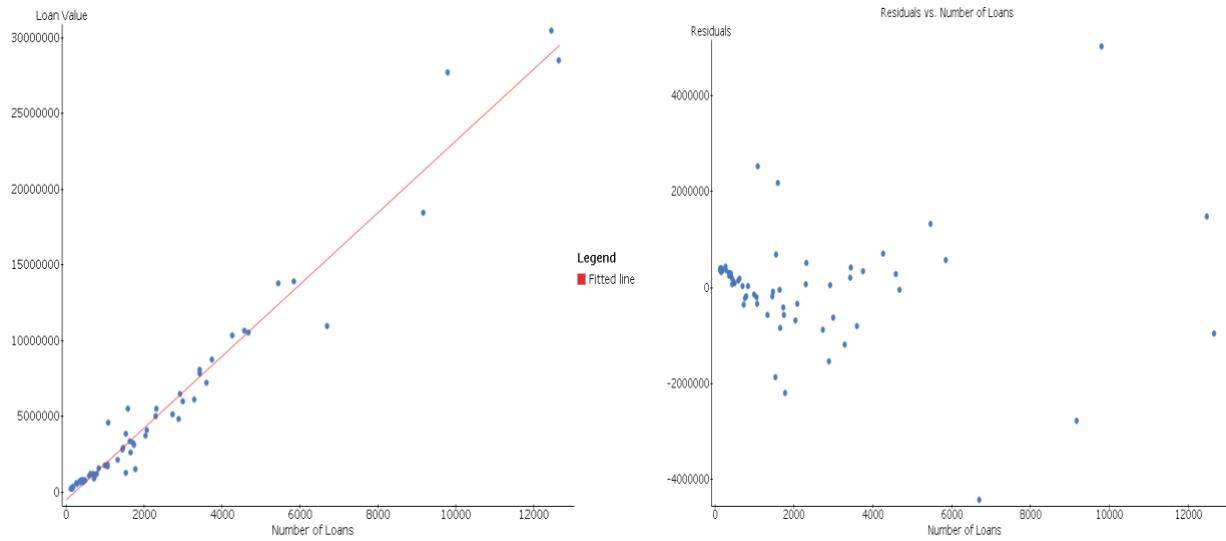


Example: Let's look at fitting a linear equation to the “**Mold Colonies**” dataset. We would like to create a linear equation to predict the “**Mold Count**” based on its “**Hours**”.

Comment below on the conditions needed to proceed with a linear regression analysis.



Example: Let's look at the scatterplot and a residual plot for the entire “**Direct Loans**” dataset.



Blackboard Video 9-J

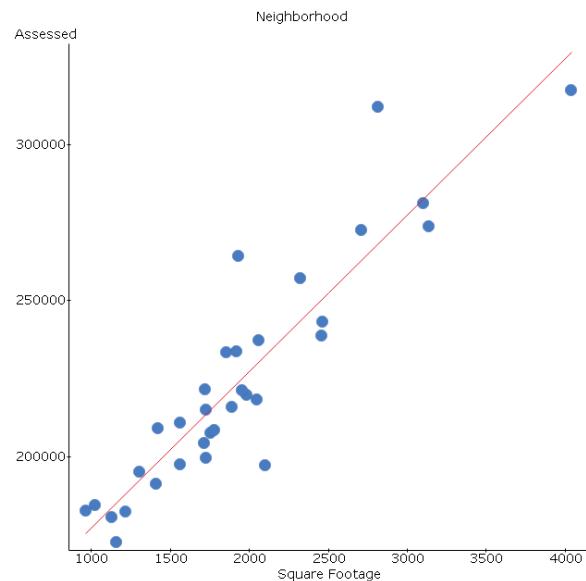
Using Residuals To Estimate the Quality of the Regression Equation

- Every data point in our regression analysis will have a residual. Recall the formula:
- The mean of all the residuals will always be _____ (points above the regression line have positive residuals and points below the line have negative residuals).
- The _____ in the residuals is the key to assessing your model.
- We would like to measure the variation in the residuals by estimating the standard deviation (StatCrunch will calculate this):
- The absolute smallest S_e can be is _____. This means there is no variation of data points about the linear regression equation. The data points will fall _____ on the line.
- If the data points are tightly packed about the regression line, then S_e will be _____. In these cases, we will have more confidence when using the equation to make predictions.
- If the data points are very scattered about the regression line, then S_e will be _____. In these cases, we will have less confidence when using the equation to make predictions.

- We cannot look at S_e in isolation; we must put its value into the context of the dataset we are analyzing.
- To do this, we will compare the variation in the residuals, S_e to the variation in the response variable, y , which we label S_y .

Example: For the “Neighborhood” dataset, we are using x = “**Square Footage**” to predict y = “**Assessed**” value. Below are the relevant standard deviations along with the scatterplot:

	Standard Deviation Symbol	Standard Deviation Value
“ Assessed ”	S_y	\$36,736
Residuals	S_e	\$15,101



Put into words what the value of \$36,736 means in the context of the problem:

Put into words what the value of \$15,101 means in the context of the problem:

- If the relationship between “**Square Footage**” and “**Assessed**” value was perfect, then $S_e = \underline{\hspace{2cm}}$. This means our x variable “**Square Footage**” perfectly predicts our y variable “**Assessed**” value.
 - If our linear equation was entirely worthless, then $S_e = \underline{\hspace{2cm}}$. This means our x variable “**Square Footage**” has no linear predictive value for our y variable “**Assessed**” value.
 - Obviously, in nearly all real-world, real-data problems, we will be somewhere in between these two extremes. What we do is look at the ratio of these two statistics to gain understanding on how well the linear equation fits our data.
 - So out of a total amount of variation of $S_y = \$36,736$ for our y variable “**Assessed**” value, the linear equation **left out** the variation in the residuals of $S_e = \$15,101$.
 - That means that $\$36,736 - \$15,101 = \$21,635$ of the variation in the y -variable “**Assessed**” value **was** accounted for by the linear equation.
 - **FYI**, in Statistics, squared units are often used (think of calculating the standard deviation and think of using Least Squares). Here, mainly for technical reasons, we will square our standard deviations – they are now called variances.
 - If we take the variance of “**Assessed**” value, we get:
-
- If we take the variance of the residuals, we get:

- Looking at the ratio of variances, we can get the percentage of the variation in our y variable ***NOT*** accounted for by the linear equation:

If the linear equation is doing a **good** job describing the linear relationship between x and y , then this percentage will be _____.

- The fraction of the variation in our y variable that **is** accounted for by our linear equation is called R^2 :

If the linear equation is doing a **good** job describing the linear relationship between x and y , then this percentage will be _____.

- For the “***Neighborhood***” dataset, we interpret R^2 as follows:

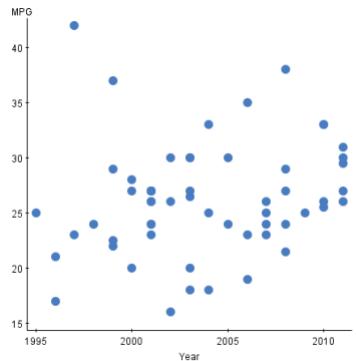
Blackboard Video 9-K

More on R-Sq

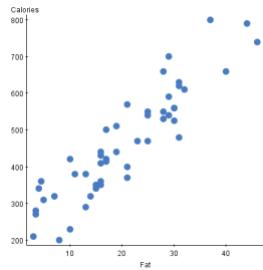
- An $R^2 = 0\%$ means that **none** of the variation in the y -variable is captured by the linear equation – the linear equation is virtually _____ . We wouldn't feel comfortable using it for prediction, and the points would be randomly scattered with no linear pattern.
- An $R^2 = 100\%$ means that all of the variation in y -variable is captured by the linear equation – our points would fall _____. We would feel very comfortable using the equation for prediction.
- What's good? It really depends on the context of the data:
 - One major company developed a method to differentiate between proteins. To do so, they had to distinguish between regressions with $R^2 = 99.99\%$ and $R^2 = 99.98\%$. For this application, $R^2 = 99.98\%$ was not good enough.
 - Separately, the president of a financial services company reports that although his regressions give $R^2 = 2\%$, they are highly successful because those used by his competitors are even lower!
- R^2 measures the _____ of variation of the y -variable accounted for by the _____ used to create the linear equation. Higher is better.
- A solid analysis of the linear relationship between two variables would include a _____, the _____, and _____.
- When interpreting the value of R^2 we must write a sentence in the context of the problem.

Example: Comment on R^2 for each model.

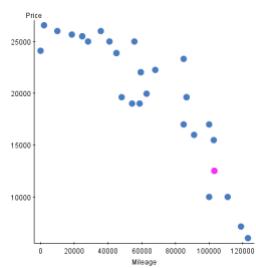
- a. Fifty-eight Math 127 students reported the “**Model Year**” and “**Miles Per Gallon**” for their primary vehicle. The linear equation had $R^2 = 3.5\%$



- b. A health food blog created a linear equation to predict “**Calories**” based on “**Fat Grams**” for fifty fast-food chicken sandwiches. The analysis had $R^2 = 82.5\%$.



- c. Professor Kupe found 31 used Toyota Avalons and created a linear equation to predict “**Price**” based on “**Mileage**”. He got an $R^2 = 76.7\%$



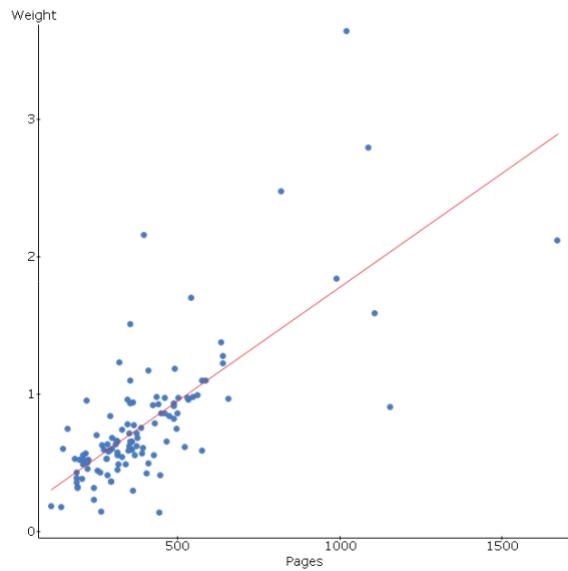
Blackboard Video 9-L

One Last Comprehensive Example

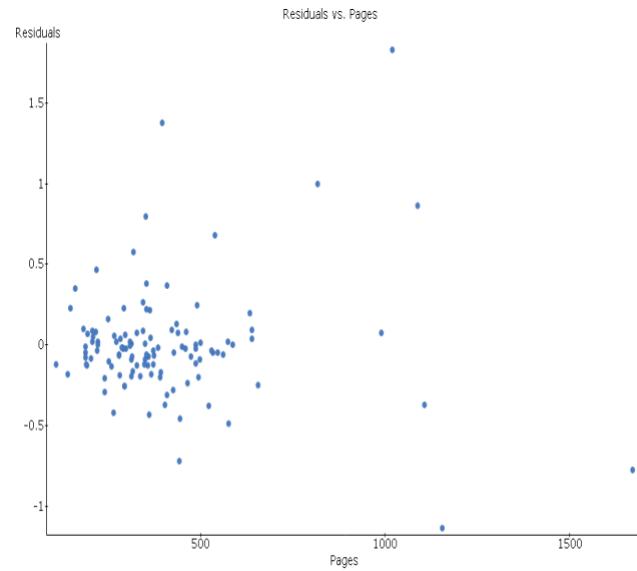
Cecil College Math 127 students visited our library a few semesters ago. Each student selected a book and the variables recorded were x = “**Number of Pages**”, used to predict y = “**Weight**” in kilograms. 115 students participated, and the relevant StatCrunch output is produced below.

Row	Book	Year	Weight	Pages	Thickness
1	Southern Women Writers	1990	0.78	344	3.5
2	Generations of Vipers	1942	0.45	313	2.5
3	Running	1973	0.626	267	2.5
4	Karl Marx : His Life and Environment	1963	0.293	361	2.1
5	Gambling	2003	0.594	358	2.4

Scatterplot



Residual Plot



StatCrunch Output

$$\text{Weight} = 0.126 + 0.00165 \text{ Pages}$$

Sample size: 115

R (correlation coefficient) = 0.729

R-sq = 53.2%

Estimate of error standard deviation: 0.355

- a. Based on the scatterplot, describe the relationship between the “*Pages*” and “*Weight*” of a book. Be sure to incorporate the correlation coefficient in your write-up.
- b. Check the conditions, using the scatterplot and the residual plot, for the appropriateness of running a linear regression analysis.
- c. Interpret, with a sentence in context, the value of the slope.
- d. Interpret, with a sentence in context, the value of the *y*-intercept.

e. For what x -values are you comfortable using the linear equation for prediction?

f. What would our equation predict for the weight of a book that has 755 pages?

g. Interpret, with a sentence in context, the value of R -sq.

h. Interpret, with a sentence in context, the value of S_e .

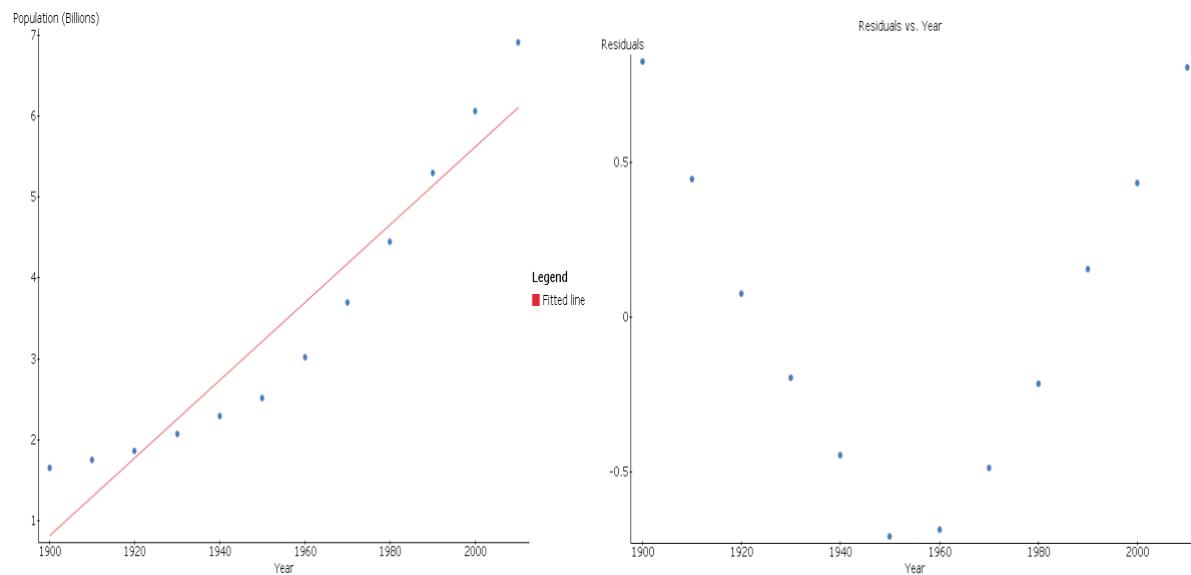
Lesson 10: Nonlinear Models

Blackboard Video 10-A

- The principle of _____ should be applied when choosing a model to fit to your data. This means that we should tend to **pick the simplest model** that does the job – in many cases, the linear model is our simplest choice – simple to understand, simple to interpret, simple to communicate to others.
- Unfortunately, **not all variables are linearly related**. It is quite inappropriate to fit a linear equation to curved data. One solution is to fit a curved model. Another is to transform your data to make it look linear. We will introduce the former in Math 127.

Example: World population is increasing over time, but not linearly. Population is increasing at an increasing rate! Here is the data, with population in billions.

1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000	2010
1.65	1.75	1.86	2.07	2.30	2.52	3.02	3.70	4.45	5.30	6.06	6.91

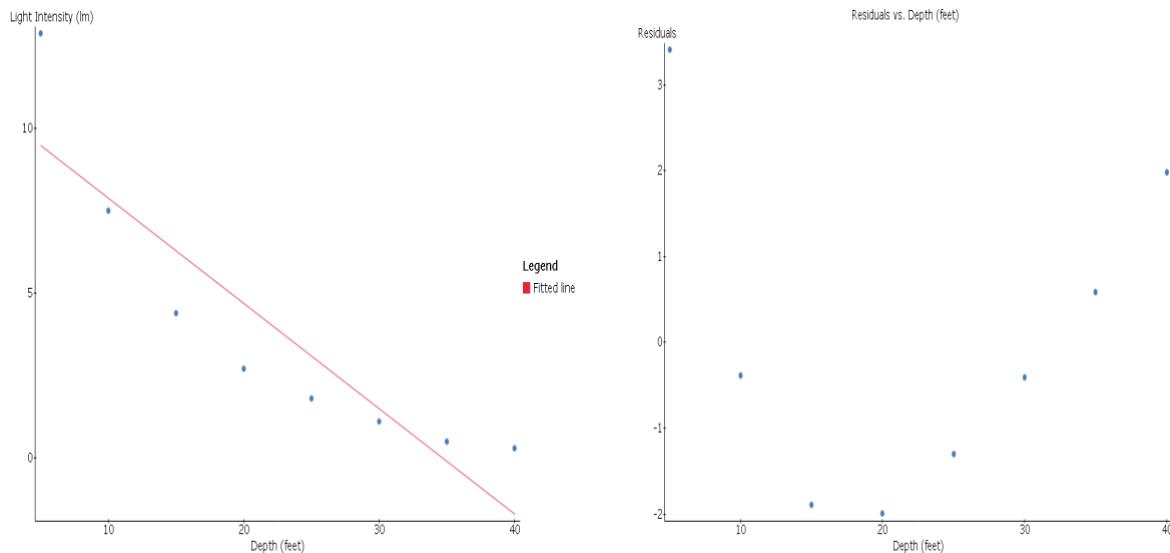


- An _____ function is good for modeling phenomena that increase at an increasing rate or decrease at a decreasing rate.
 - In general, an exponential function looks like this:
- (Graph of an exponential function)
- If the base exceeds 1, $b > 1$, then we have exponential _____.
 - If the base is between 0 and 1, $0 < b < 1$, then we have exponential _____.

Example: “**Light Intensity**” (the amount of visible light) **decreases** exponentially with “**Depth**” as we travel below the surface of a body of water. A biologist uses a photometer to measure light penetration in a Minnesota lake, obtaining the data in the table.

Depth (ft)	5	10	15	20	25	30	35	40
Light Intensity (lm)	12.9	7.5	4.4	2.7	1.8	1.1	0.5	0.3

Note: The correlation between the variables is $r = -0.898$. Without looking at a graph, you might think we have a slam-dunk linear relationship.

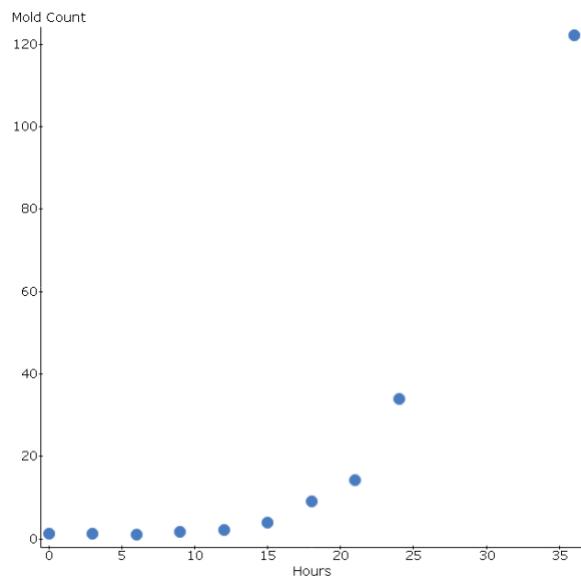


Blackboard Video 10-B

Using Exponential Functions

Example: A mold colony was cultured by a scientist in a Petri dish for one-and-a-half days. The scientist would like to use “**Hours**” elapsed to predict the “**Mold Count**” (in millions). The data, the scatterplot, and the exponential equation are given below.

Hours	0	3	6	9	12	15	18	21	24	36
Mold Count	1.23	1.18	0.94	1.7	2.13	3.99	9.02	14.27	33.89	122.26



a.

Explain in words why an exponential growth model is appropriate.

$$\text{Mold Count} = 0.6122(1.1576)^x, \quad x = \text{Hours}$$

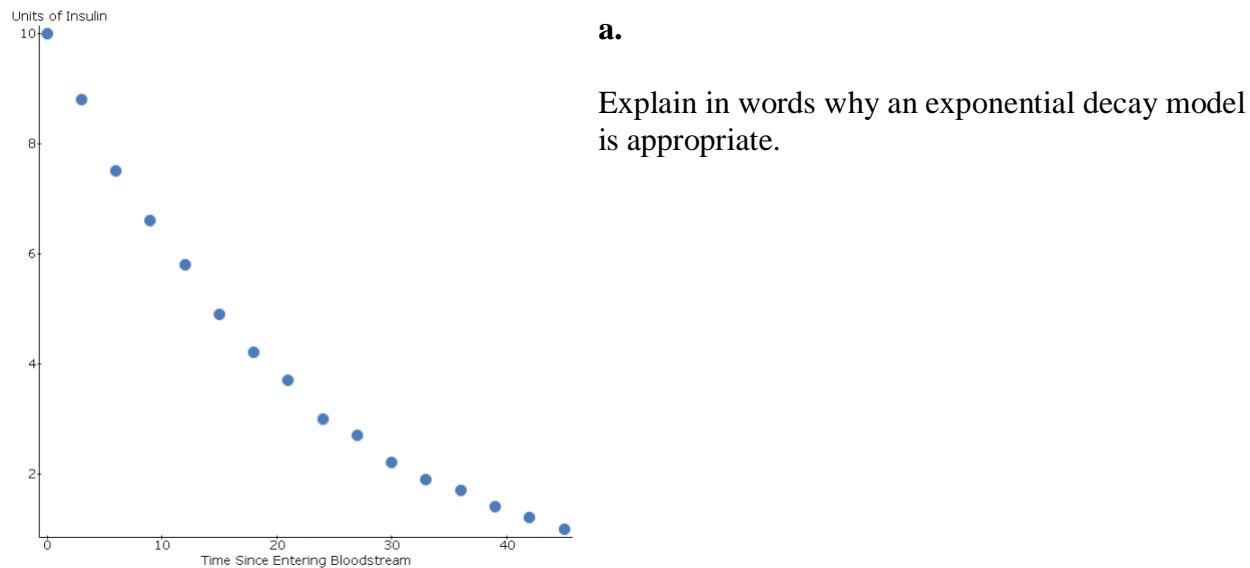
b. What would the equation predict the “**Mold Count**” to be at 30 hours?

c. Explain why we would have reservations about predicting much beyond 36 or so hours?

Example: Some percentage of the population are diabetic, in which the body is unable to produce insulin, which is needed to process glucose. To provide the hormone, an injection of medicine containing insulin is taken. As time elapses, the insulin breaks down in the body.

The variables displayed below are “**Time Since Entering Bloodstream**” in minutes and “**Units of Insulin**” in the bloodstream. A doctor took measurements on one patient in a controlled environment.

Minutes	0	3	6	9	12	15	18	21	24	27	30	33	36	39	42	45
Units of Insulin	10	8.8	7.5	6.6	5.8	4.9	4.2	3.7	3	2.7	2.2	1.9	1.7	1.4	1.2	1



- b. Which of the following equations will correctly describe the pattern in the data? Why?

$$\text{Units of Insulin} = 10.405(1.95)^x, \quad x = \text{Time}$$

$$\text{Units of Insulin} = 10.405(0.95)^x, \quad x = \text{Time}$$

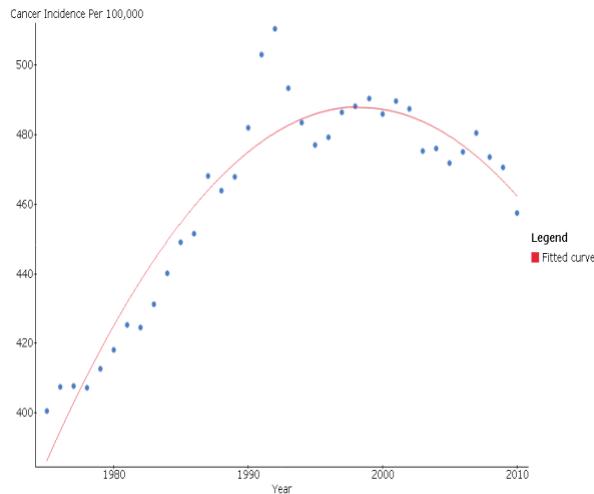
- c. Use the correct equation to predict the insulin at one hour (a little extrapolation).

Blackboard Video 10-C

Using Quadratic Functions

- The final type of model we will briefly explore is the quadratic model.
- Quadratic data is fairly uncommon in a Statistics setting, but if you encounter it, look for that familiar _____ shape.
- In general, a quadratic equation looks like this:

Example: The “*Number of Cancer Incidences*” per 100,000 Americans was recorded from 1975 through 2010 by a government agency. The scatterplot and quadratic equation are below.

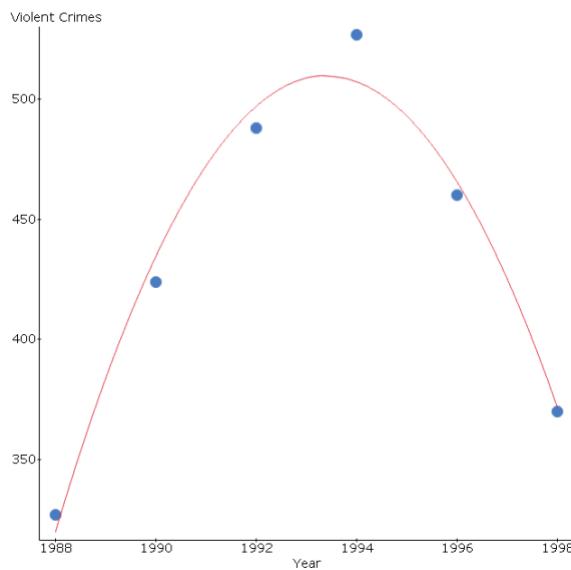


Note: It is probably most-dangerous to extrapolate when using a quadratic equation for real-world data. Just because the pattern for the past 35 years can be somewhat described by a parabola, it does not mean cancer rates will continue to decline, even into next year.

$$\text{Cancer Incidences} = -0.1870476(\text{Year})^2 + 747.55737(\text{Year}) - 746,437$$

- a. In 1992, the actual cancer incidence peaked at 510.6 per 100,000 Americans. What does the quadratic equation predict and then give the value of the residual.

Example: USA Today reported the “*Number of Juvenile Violent-Crime Arrests*” by “*Year*” for about a decade. The scatterplot and quadratic equation are below:



$$\text{Violent Crimes} = -6.513393(\text{Year})^2 + 25,967.555(\text{Year}) - 25,881,314$$

- a. Use the equation to predict violent crimes in 1997. Do we have any good reason to have faith in our prediction?

Concluding Remarks

- When choosing between a linear, exponential, and quadratic, you must always _____ first, to see which model is most appropriate.
- If a linear model is appropriate, we can measure the fit of the model by looking at _____. Higher is better, which means your model has more _____ value.
- Not all data will lend itself to being modeled – use your best judgment and remember to apply the principle of _____. Simpler is usually better.
- Finally, all models are _____. We hope to find one that is at least _____.

Lesson 11: Probability

Blackboard Video 11-A

Why Study Probability?

- There are two sides to the field of Statistics – _____ statistics and _____ statistics. In a way, probability is a link between the two. Draw a diagram below:

Definitions (Using Math 127 attendance at Cecil College as an example)

- A _____ is a repeatable action or observation that has an uncertain result.
- For a Math 127 class, the number of student that actually show up on any given day is a _____. We won't know how many show up tomorrow until we actually hold class and count up the number of attendees.
- A full Math 127 class has 24 students, and for simplicity, assume we are dealing with a full class. Your instructor is interested in days where **at least 22 student show up**; she considers that a success. This is called an _____ in our probability experiment. All 24 students showing up would be another event.
- Each class meeting is a _____ in this **probability experiment**, and each class, whether or not at least 22 students show up, is called an _____ in our probability experiment.
- One way to estimate probability is by collecting data. This is called the _____ method to determine probability.
- Over time, as more and more Math 127 classes are held, results to this probability experiment accumulate. Your instructor keeps a running tally of the number of classes that at least 22 students show up. If taken as a _____, this proportion or percentage will be used to estimate the probability of 22+ students attending a Math 127 class. For example:

- One assumption required for most every probability experiment is that the trials are _____ from each other. In other words, today's attendance has no impact on tomorrow's attendance. We will assume so here.
- The _____ states that as we complete more trials, our relative frequency estimate of the true probability should tend to get closer to the true value. In other words, when estimating probability empirically, **more data will give us better estimates**. For our example, as we hold more Math 127 classes, our estimate of the probability of 22+ attendees will become more accurate and more reliable.
- A **second way** to **estimate** the **true probability** of an event is to use _____ probability. This method is **not** based on data, but rather personal, subjective belief of the likelihood of an uncertain outcome.
- For example, if the Dean at Cecil College asked the Math 127 instructor, “What is the probability a Math 127 class has at least 22 students in attendance?”, and off-the-cuff, the instructor replied, “About 80%”, the instructor is using personal experience to give a best guess of the true probability. This is **subjective**, as no data was used to support the estimate.
- A **third way** to **estimate** the **true probability** of an event is called the _____ method. Using this method, we assume all simple events are equally likely to occur.
- The classical method would fail for estimating Math 127 class attendance because it assumes every outcome is equally-likely. In other words, it is just as likely to have 22 students attend as it is to have 3 students attend. As the instructors, we certainly hope this isn't the case.
- The classical method **would** work for the probabilities associated with rolling a fair die:

Blackboard Video 11-B

Two Probability Distributions

- One way to describe a probability distribution is to list the individual outcomes along with the associated probabilities.

Example: For a class of 24 Math 127 students, a probability distribution is given for the number of students who show up on any given day. The instructor used subjective probability to derive this table.

# of Students	24	23	22	21	20	19	18	17	16	15 and under
Probability	0.10	0.20	0.20	0.15	0.10	0.08	0.07	0.04	0.03	0.03

- Give the probability that 1 or 2 students are absent.
- What's the probability at most one student is absent, based on this distribution?
- What's the probability that at least 4 students are absent?
- What is the sum of all the simple probabilities for this probability distribution?

- A second way to describe a probability distribution is to give a mathematical function and associate the area under the curve with the probability of certain outcomes.

Example: The train you take home from work each day in NYC arrives like clockwork every 10 minutes at your stop. The time you get to the platform each day is more or less completely random because you leave work at different times each day. Your “*Wait Time*” for the train can be described by a **uniform** probability distribution because every imaginable wait time is equally likely.

a. What is the shortest and longest time periods you could wait for the train?

b. What is your average wait time?

c. Draw a graph of this probability model.

- d. What is the probability that we wait less than two minutes for the train? Draw and shade the graph.
- e. What is the probability that we wait at least 6.5 minutes for the train? Draw and shade the graph.
- f. Draw, label, and interpret the 90th percentile for this probability distribution.

Blackboard Video 11-C

Probability Rules and Definitions Part 1

Rule 1. Valid Probability Values

Example: For the Math 127 attendance distribution, notice all probabilities are legal values.

# of Students	24	23	22	21	20	19	18	17	16	15 and under
Probability	0.10	0.20	0.20	0.15	0.10	0.08	0.07	0.04	0.03	0.03

Rule 2. The sample space has a probability of 1

Example: Verify that the Math 127 attendance distribution follows Rule 2.

Example: You subjectively create the following table for the chances you graduate in the next few years. What is the probability that you drop out of college?

Graduate in...	2 Years	2.5 Years	3 Years	3.5 Years	Drop Out
Probability	0.15	0.20	0.50	0.10	

Rule 3. The Complement Rule

Definition: The complement of event A is the set of outcomes that is not in event A .

Example: The probability that the Browns beat the Ravens in Cleveland is 0.21. What is the complement and what is its probability?

Example: For the Math 127 attendance distribution, determine the following complements and probabilities.

# of Students	24	23	22	21	20	19	18	17	16	15 and under
Probability	0.10	0.20	0.20	0.15	0.10	0.08	0.07	0.04	0.03	0.03

a. What is the complement of “*Perfect Attendance*”? Give its probability.

b. What is the complement of “*At Least 22 Students Attend*”? Give its probability.

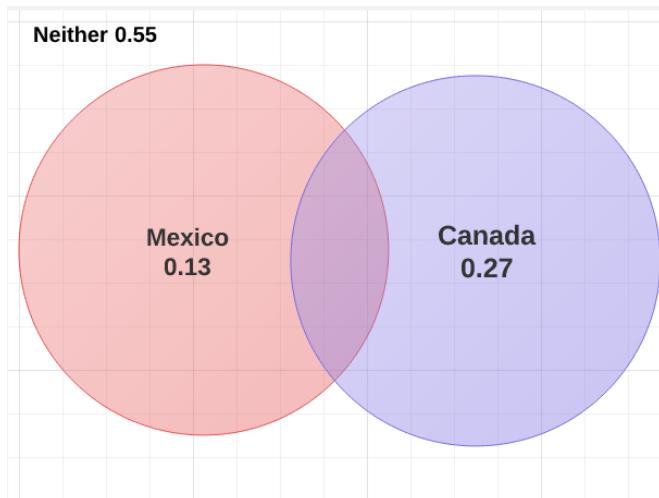
Blackboard Video 11-D

Rule 4. Addition Rule for Disjoint Events

Definition: Events are disjoint if they cannot occur at the same time. Mutually exclusive is synonymous with disjoint.

Example: In the Math 127 attendance example, all the outcomes are disjoint. We cannot have 24 and 23 students show up on the same day. What is the probability that 17 or fewer students attend?

Example: Below is a Venn Diagram showing the probabilities that a Cecil College student has been to Canada or Mexico.



c. What is the probability that a student has been to Canada?

- Are the events "*Mexico*" and "*Canada*" disjoint?
- What is the probability that a student has been to both Mexico and Canada?

Rule 5. Multiplication Rule for Independent Events

Definition: Two events are **independent** if the occurrence of one event **does not** influence the probability of the other event happening.

Example: If we flip a fair coin multiple times, each flip is independent of any previous flips. No matter what the outcome has been on your first few flips,
 $P(\text{Head}) = \underline{\hspace{2cm}}$ and $P(\text{Tail}) = \underline{\hspace{2cm}}$.

a. Give the probability of flipping a head two times in a row.

b. Give the probability of flipping a tail 5 times in a row.

Example: At Cecil College, roughly $\frac{2}{3}$ of all students are female. If we randomly select 4 students to represent the college at a retreat, what is the probability that the group is entirely female?

Blackboard Video 11-E

Example: It is known that approximately 45% of all people have Type O blood, 40% Type A, 11% Type B, and 4% Type AB. No one can have more than one type of blood.

a. Draw and label a Venn diagram.

b. If we randomly select one person at the blood drive, give the probability that they are Type A or Type AB?

c. Give the probability that a random adult is *not* Type O?

- d. Presuming the next two people that walk through the door are unrelated, what is the probability they are **both** Type A?
- e. Suppose we desperately need some Type A blood. What is the probability that for the next two donors, **at least one of them** is Type A?

Example: At Cecil College, it is known that 85% of Math 127 students have never taken any calculus, 10% have taken only Calculus I, and 5% have taken two or more calculus courses.

- a. Your instructor is going to create random groups of four students. Give the probability that no one in the group has ever studied any calculus.

- b. All four members of the group have had some calculus?

- c. In your group of four, give the probability that at least one person in the group has studied some calculus?

Blackboard Video 11-F

Probability Rules and Definitions Part 2

- It would be nice if every pair of events were disjoint, but in reality, many events are not disjoint. We need an addition rule to accommodate this situation.

Rule 6: The Addition Rule for Non-Disjoint Events

Example: All Cecil College graduates must take one college-level math course for their degrees, and most students will take either Precalculus or Statistics. The $P(\text{Precalculus}) = 0.23$ and $P(\text{Statistics}) = 0.71$. We also know that 15% of our graduates have taken both courses.

a. Draw a Venn diagram of the situation.

b. Give the probability that a student has taken Precalculus or Statistics. (If said this way, it is always implied that the student has taken Precalculus, or Statistics, or both courses).

Example: A randomly selected Cecil student has a 74% probability of having only a Facebook account, a 6% probability of having only a Twitter account and a 4% probability of having neither.

Blackboard Video 11-G

Conditional Probabilities

- The probability of event B occurring if we know event A has already occurred is a **conditional** probability.
- We denote it as:

Rule 7: The formula to compute a conditional probability is given as:

Example: At Cecil College, the following data was collected. Compute the following probabilities.

	Male	Female	Total
Nursing Student	3	29	32
Not a Nursing Student	129	239	368
Total	132	268	400

a. $P(\text{Male} \mid \text{Nursing Student}) =$

b. $P(\text{Male} \mid \text{Not a Nursing Student}) =$

c. $P(\text{Male}) =$

d. $P(\text{Nursing Student} \mid \text{Male}) =$

e. $P(\text{Nursing Student} \mid \text{Female}) =$

f. Comment on whether or not “*Gender*” and “*Are You a Nursing Student*” are independent. Support with probabilities.

Blackboard Video 11-H

- It would be nice if every pair of events were independent, but in reality, many events are dependent. We need a multiplication rule to accommodate this situation.

Rule 8: The Multiplication Rule for Dependent Events

A Test for Independent Events

- Events A and B are independent if:

A Test for Disjoint Events

- Events A and B are disjoint if:

All This In Words

- Events are _____ if the occurrence of one event _____ change the probability of the next event.
- Events are _____ if the two events cannot occur at the same time.
- Disjoint events _____ be independent, since if event A occurs, it is guaranteed that event B _____ occur.

Example: We know $P(\text{Head}) = 0.5$ for a fair coin.
What is the $P(\text{Head on 2}^{\text{nd}} \text{ flip} \mid \text{Head on 1}^{\text{st}} \text{ flip})$?

Example: We know that $P(\text{Female}) = 0.62$ at a certain college.
Also, the $P(\text{Engineering Major} \mid \text{Female}) = 0.03$. What is the probability that a randomly selected student is a female engineering major?

Example: You can either walk or take the bus to class. $P(\text{Late for Class} \mid \text{Took the Bus}) = 0.35$ and $P(\text{Late for Class}) = 0.20$. Are “**Mode of Transportation**” and “**Whether or Not You’re Late**” independent or dependent? Why?

Example: Same as above. $P(\text{Late for Class} \mid \text{Took the Bus}) = 0.35$ and now we know $P(\text{Took the Bus}) = 0.50$. On any random day, what is the probability that you were late to class *and* took the bus?

Blackboard Video 11-I

Example: In 2008, 49% of small employers offered health insurance, 14% offered a retirement package, and 9% offered both benefits. We will choose a small employer at random.

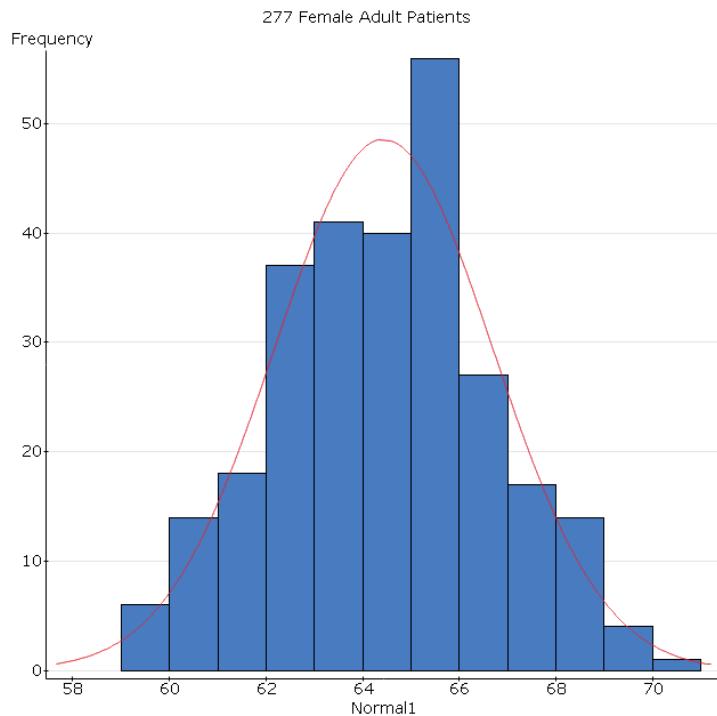
- e. Draw a Venn diagram for this scenario.
- f. Show, in two different ways, that offering “***Health Insurance***” and offering a “***Retirement***” package are dependent upon each other.

Lesson 12: The Normal Model

Blackboard Video 12-A

- The **Normal model** is a good probability model for quantitative variables when the histogram looks _____ and _____.
- A **Normally distributed** variable has the familiar _____ curve.
- The Normal model is very important for two reasons. The **first** reason: many variables in practice are well-modeled by a Normal curve. Examples include heights of males and females (separately), weights and other measurements of things that are manufactured, IQ scores, newborn birth weights, and a plethora of other natural phenomenon.
- If you collect data on a variable you are analyzing, you must graph your data with a _____ to look for that bell-shaped distribution. A second graph we will check, called a _____, will be introduced in a few pages. Do **not** assume a variable is Normal without checking the graphs first.

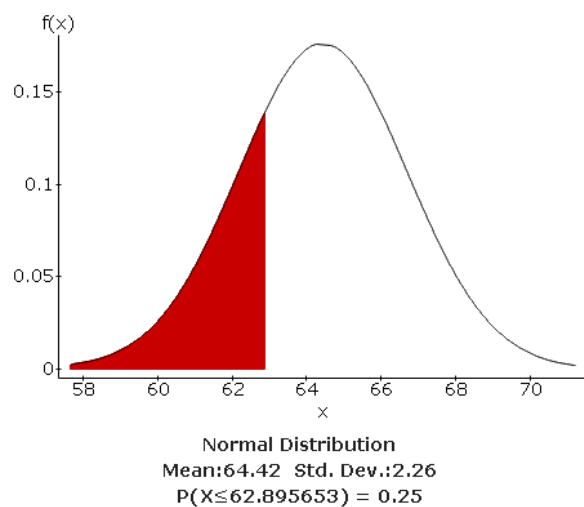
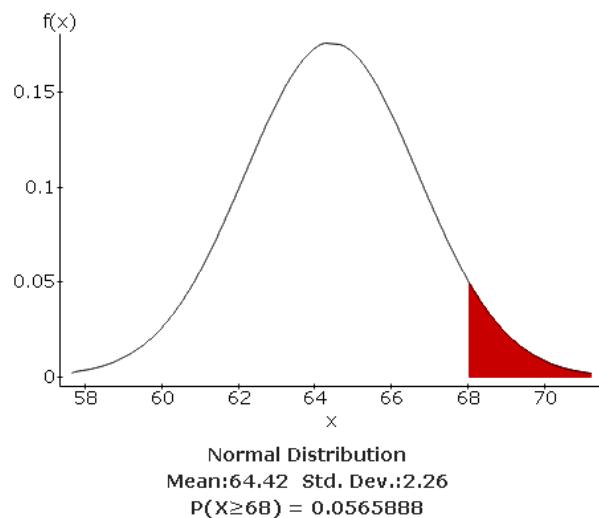
Example: A Baltimore doctor graphed the heights (inches) of his 277 female patients.
Comment.



- The **second** reason Normal models are so important is that theoretically, many statistics like the sample proportion (Lesson 14) and the sample mean (Lesson 18) will follow a Normal model, under certain conditions. Because of this amazing fact, we can build the foundation of all of our methods of inference in Math 127 (Unit III).
- A **Normal model** is a **probability model**. We can find associated probabilities by finding the area under appropriate Normal model.

Example: The doctor from the previous page used the appropriate Normal model to answer the following questions:

- What percentage of the female patients in Baltimore are 5' 8" or taller?
- What is the 25th percentile for female height?



- Every Normal model is identified by two parameters –
 - The _____, which for models is called _____.
 - The _____, which for models is called _____.
- By the way, a parameter is a number we choose to specify a model. We don't want to confuse parameters with statistics – recall when we collect data, we can summarize it by calculating summary statistics:
 - Remember, after collecting data, we can calculate the mean and call it _____.
 - We can also calculate the standard deviation from the data and call it _____.
- The shorthand notation to describe a **Normal model** is: _____. This is how we say, "Here is a Normal model with a mean of μ and a standard deviation of σ ".

Example: IQ scores are Normally distributed with a mean of 100 points and a standard deviation of 15 points.

- a. What is the shorthand notation for this Normal model? _____
- b. Draw the correct Normal model and shade it to represent the probability that a randomly selected adult has above average IQ.

Blackboard Video 12-B

The Empirical Rule / Three-Sigma Rule

- Every Normal model follows what is called the **empirical rule** or the **three-sigma rule**.
To make the numbers cleaner and easier to remember, we will do a bit of rounding.
- About _____ of the values lie within _____ standard deviation of the mean.
- About _____ of the values lie within _____ standard deviations of the mean.
- About _____ of the values lie within _____ standard deviations of the mean.
- This leaves about _____ of the values split equally between the lower and upper tails remaining.

Draw a Normal model labeled with the empirical rule:

Draw a second Normal model, partitioned and labeled using the empirical rule:

- Remember we've stated that values within _____ standard deviations of the mean are not unusual. For Normal models, this corresponds to the central _____ of all values.
- We've stated that values between _____ and _____ standard deviations away from the mean are "**unusual**".
- We've stated that values exceeding _____ standard deviations are "**rare**". From a probability standpoint, you can now see why.

Example: Draw a partitioned Normal model for IQ scores using $N(100,15)$.

- a. What percentage of people have IQs exceeding 130? _____
- b. What percentage of people have IQs between 85 and 115? _____
- c. What IQ score represents the 16th percentile? _____
- d. What IQ scores are not unusual? _____
- e. What IQ scores are unusual? _____
- f. What IQ scores are rare? _____

Blackboard Video 12-C

Linking Z-Scores to the Normal Model

- Remember from previous lessons that a z -score is the number of standard deviations a value lies away from the mean. This idea translates nicely when working with Normal models.
- Recall the z -score formula when you are working with data:
- Here is the z -score formula for when you are working with a Normal model (any model):

Example: IQ scores are well-modeled by $N(100,15)$.

- a. What is the z -score for someone with an IQ of 83? Is that unusually below average?
- b. Your girlfriend has an IQ with a z -score of 2.80 and claims to be a genius. Is she? What's her IQ?

The Standard Normal Model

- There are an **infinite** number of Normal models. Because Normal models are defined by the value of their mean (which could be anything) and their standard deviation (which could be anything positive), there are unlimited possibilities.
 - General population IQ scores follow $N(100, 15)$, which is centered at 100 and has a spread of 15 IQ points. Maybe at Harvard, $N(105, 10)$ describes the distribution of IQ scores. We'd have to collect data to find out.
 - In America, women's heights can be modeled with $N(64, 2.3)$ while men's heights can be modeled with $N(70, 2.5)$. But in Japan, perhaps different Normal models are appropriate.
- One very important Normal model is the _____.
- One way to think about the Standard Normal model is that it is the Normal model for _____. Any Normal model can be _____.
- Take, for example, the IQ Normal model and convert the values to z -scores. You have now converted $N(100, 15)$ to $N(0, 1)$, which is the Standard Normal model.
- For **any dataset** or **any model**, the mean of the z -scores is always _____. Therefore the mean of the Standard Normal model is always _____.
- For **any dataset** or **any model**, the standard deviation of the z -scores is always _____. Therefore the standard deviation of the Standard Normal model is always _____.

Example: Draw a partitioned Standard Normal model. Label with the empirical rule.

Blackboard Video 12-D

Example: A company that manufacturers rivets believes the shear strength (in pounds) is well-modeled by $N(800,50)$.

If we're building a bridge requiring thousands of these rivets, would it be safe to use these rivets in a situation requiring sheer strength of 750 pounds? Explain.

What if we found a competing supplier of rivets who claims their rivets follow a $N(795,15)$ model? Should we use this supplier?

Example: A winemaker has two cork cutting machines at the winery. Machine A cuts corks with diameters that follow $N(3, 0.1)$ in centimeters, while Machine B follows $N(3.04, 0.02)$.

Blackboard Video 12-E

Using StatCrunch to Solve Normal Model Problems – Part 1

- Most times, we will be answering Normal model questions that do not fall exactly on 1, 2, or 3 standard deviations away from the mean. In these cases, the 68-95-99.7% empirical rule will not be sufficient.

Example: Based on the Normal model for IQ, $N(100, 15)$, use StatCrunch and draw pictures to answer the following questions.

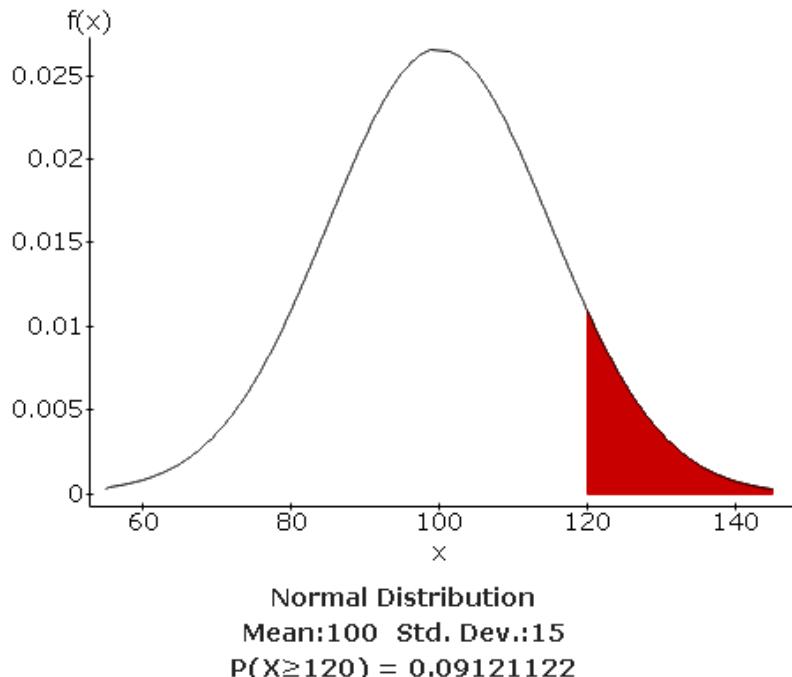
- a. There are many ways to essentially ask the same question. Give the probability that a randomly selected person has an IQ exceeding 120. Or, what percentage, fraction, proportion of the population has an IQ score of 120 or above.

From any StatCrunch worksheet (blank or a dataset):

StatCrunch \rightarrow Stat \rightarrow Calculators \rightarrow Normal

Type In: Mean=100 Std. Dev. = 15

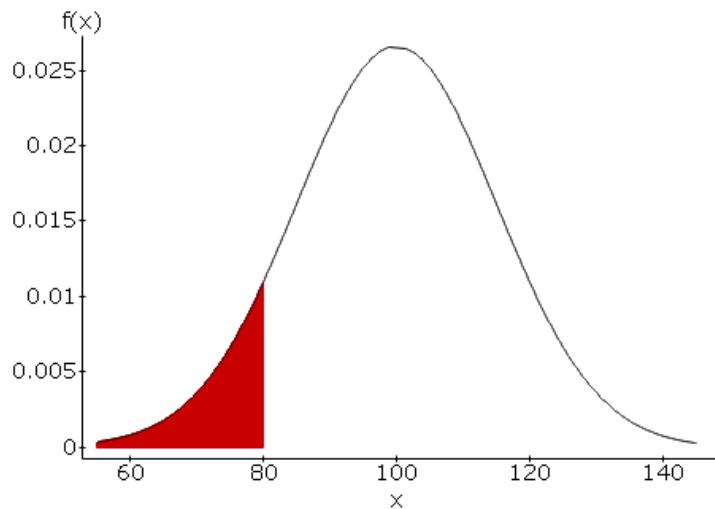
We Want: $P(X \geq 120) =$ _____ to answer the above question. Hit Compute.



- b. What is the probability that someone in Maryland has an IQ of at most 80?

Type In: Mean = 100 Std. Dev. = 15

We Want: $P(X \leq 80) = \underline{\hspace{2cm}}$ to answer the above question. Hit Compute.



Normal Distribution
Mean: 100 Std. Dev.: 15
 $P(X \leq 80) = 0.09121122$

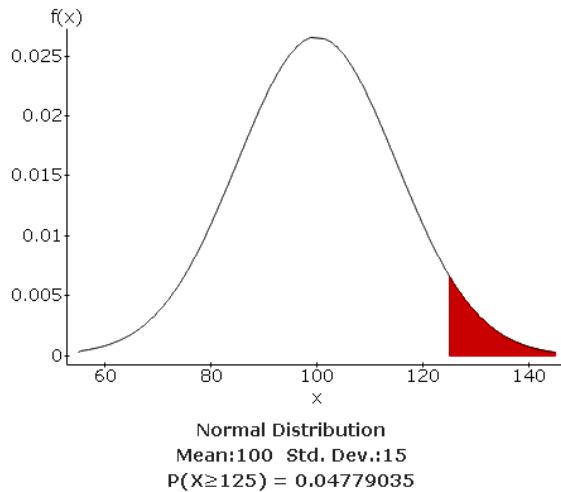
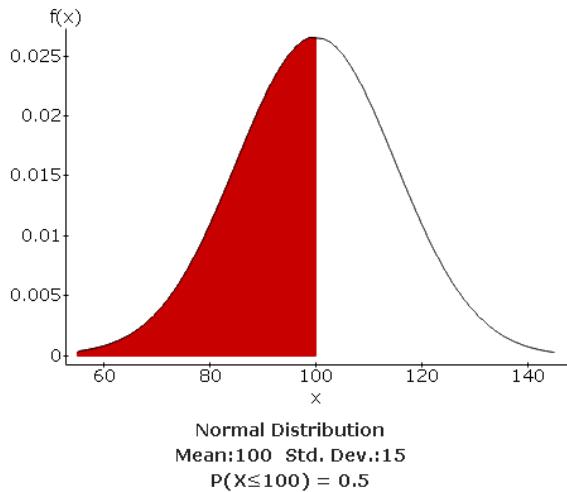
- c. What percentage of the population have IQ scores that are at least 3.5 standard deviations above the mean? Use StatCrunch.

In-Betweeners

- d. Give $P(100 \leq X \leq 125)$ using StatCrunch.

Because StatCrunch will not compute an “in-between” area under a Normal model, we can use the following logic to arrive at the answer.

1. Find the area to the left of the left boundary. Write it down.
2. Find the area to the right of the right boundary. Write it down.
3. Because the entire area is 100% = 1, take $1 - (\text{area to the left}) - (\text{area to the right})$.



Now Practice

- e. Percentage of people with IQs between 80 and 90? _____
- f. Probability the person sitting next to you in class has an IQ of at most 95? _____
- g. Proportion of Cecil College students with IQs greater than 150? _____
- h. If there are 2500 Cecil students, how many do we expect to have IQs of at least 150?

Blackboard Video 12-F

Using StatCrunch to Solve Normal Model Problems – Part 2

- In Video 12-E, we solved problems where we knew the boundaries on the horizontal axis and we were looking for the area to the left, to the right, or in-between.
- Now, we will know the area breakdown and we will be solving for the values on the horizontal axis that led to that area breakdown. Essentially, we're going in reverse now.

Example: Based on the Normal model for IQ, $N(100, 15)$, use StatCrunch and draw pictures to answer the following questions.

- Give the IQ score that marks the 30th percentile of all IQ scores.

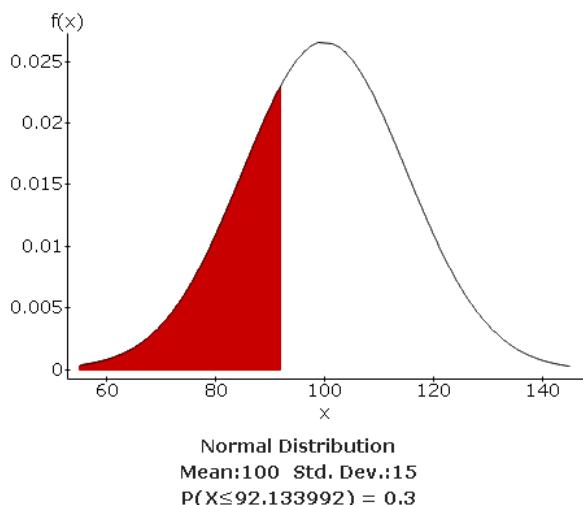
Put in words what the “30th percentile” means:

From any StatCrunch worksheet (blank or a dataset):

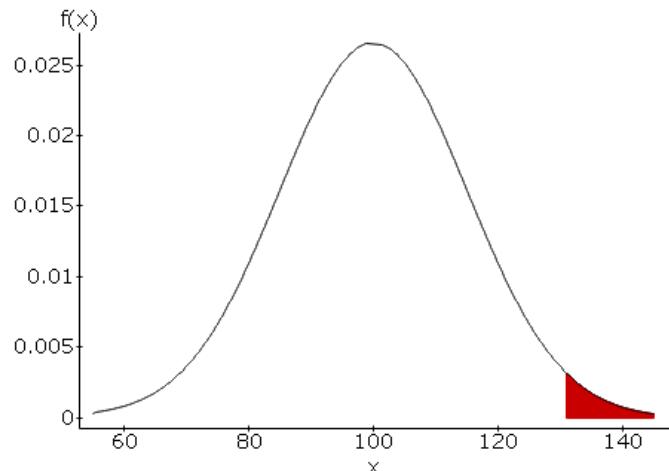
StatCrunch → Stat → Calculators → Normal

Type In: Mean = 100 Std. Dev. = 15

We Want: $P(X \leq \text{_____}) = 0.30$ to answer the above question. Hit Compute.

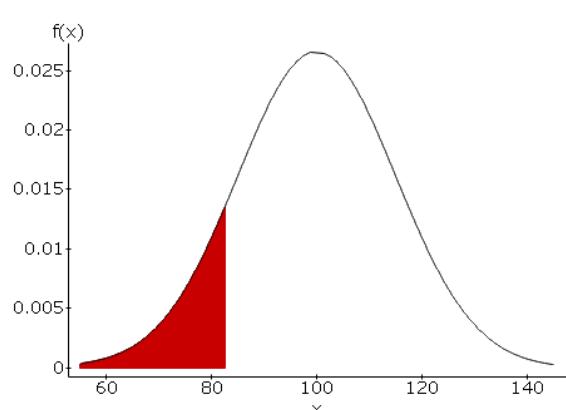


- j. What IQ score separates the top 2% of people from the rest of us?

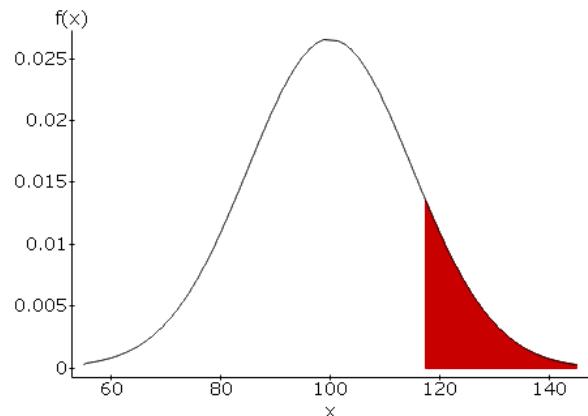


Normal Distribution
Mean:100 Std. Dev.:15
 $P(X \geq 130.80623) = 0.02$

- k. Between what two IQ scores do we find the central 75% of all people?



Normal Distribution
Mean:100 Std. Dev.:15
 $P(X \leq 82.744759) = 0.125$



Normal Distribution
Mean:100 Std. Dev.:15
 $P(X \geq 117.25524) = 0.125$

Blackboard Video 12-G

Example: Men's heights are often modeled with $N(70, 2.5)$.

a. What percentage of men exceed 6 feet tall? Use StatCrunch but draw in the picture:

b. Convert 72 inches to a z -score. Draw and shade the Standard Normal model:

- c. We know this guy who has a z -score of -1.38 . At what percentile is his height. What is his height? Draw Normal models as needed.
- d. The shortest 3% of all grown men stand how tall (or shorter)? What is the z -score that goes with this height?

Blackboard Video 12-H

Example: Grocery shoppers likely bypass any apples that weigh less than 80 grams. The current genetically-modified Granny Smith plant produces fruits that average 100 grams, but 10% of the apples are too small. It is reasonable to assume a Normal model applies.

a. What is the standard deviation of the weights of the apples currently being grown?

b. For these apples, give the 90th percentile and explain what it means.

- c. Suppose we know that Fuji apples also follow a Normal distribution, but this time the mean and standard deviation are both unknown. Consumers still bypass any apples weighing less than 80 grams, and for Fuji apples, this happens for 15% of all apples.

Additionally, for Fuji apples, the heaviest 5% weigh at least 105 grams. Find the mean and standard deviation for the Fuji apple Normal model.

- d. McIntosh apples have a 25% chance of being bypassed by consumers (under 80 grams). Only 2% exceed 100 grams. A Normal model applies; find the mean and standard deviation.

Blackboard Video 12-I

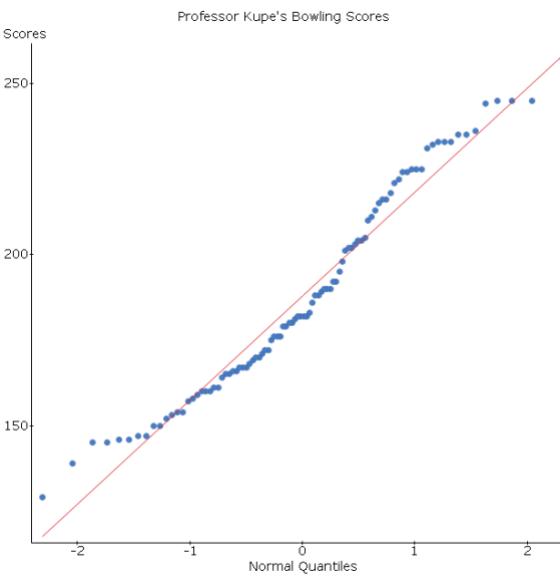
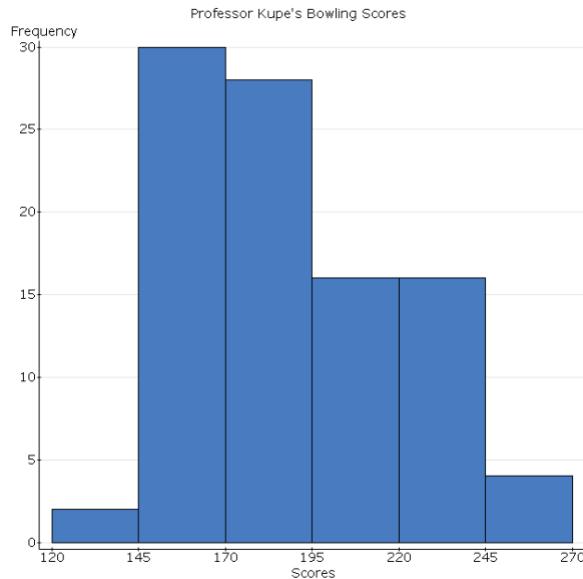
Checking Data for Normality

- It may be tempting to assume that a Normal model always applies to your data. This is not the case!

How to use StatCrunch to determine if a Normal model is appropriate for a variable:

1. Make a histogram. The data must be _____, _____, and have minimal outliers.
2. Make a QQ plot. QQ stands for _____. In this plot, if the points are generally straight and generally fall close to the diagonal line, then normality is plausible³. A QQ plot is not a scatterplot, so don't interpret it as such.

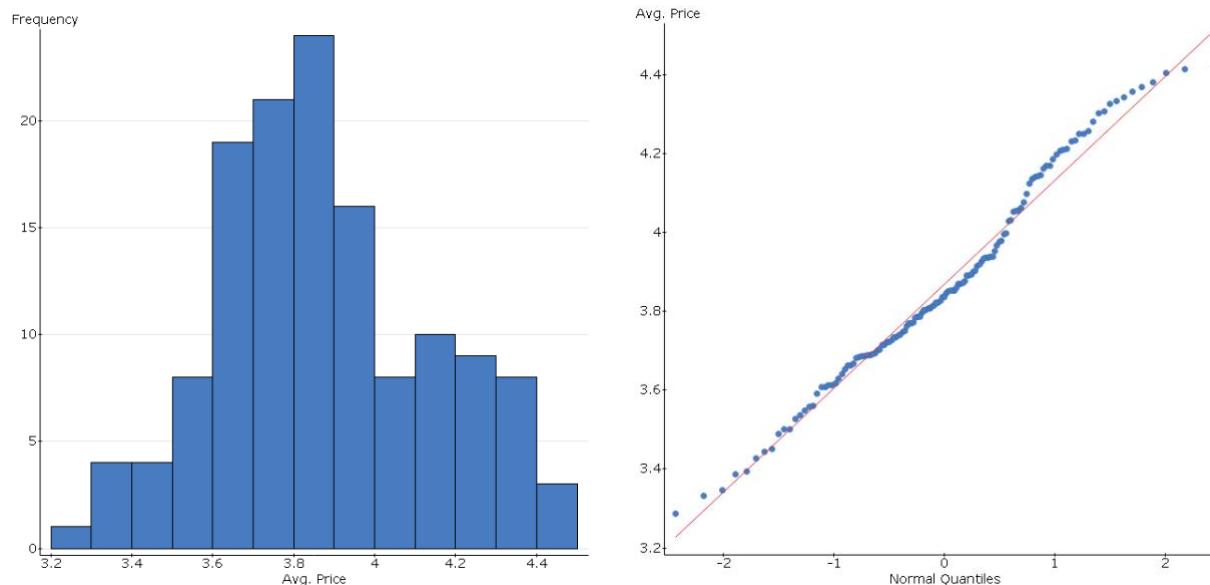
Example: A few seasons ago, Professor Kupe bowled in Men's League and kept track of his scores. Make the call, Normal or not Normal.



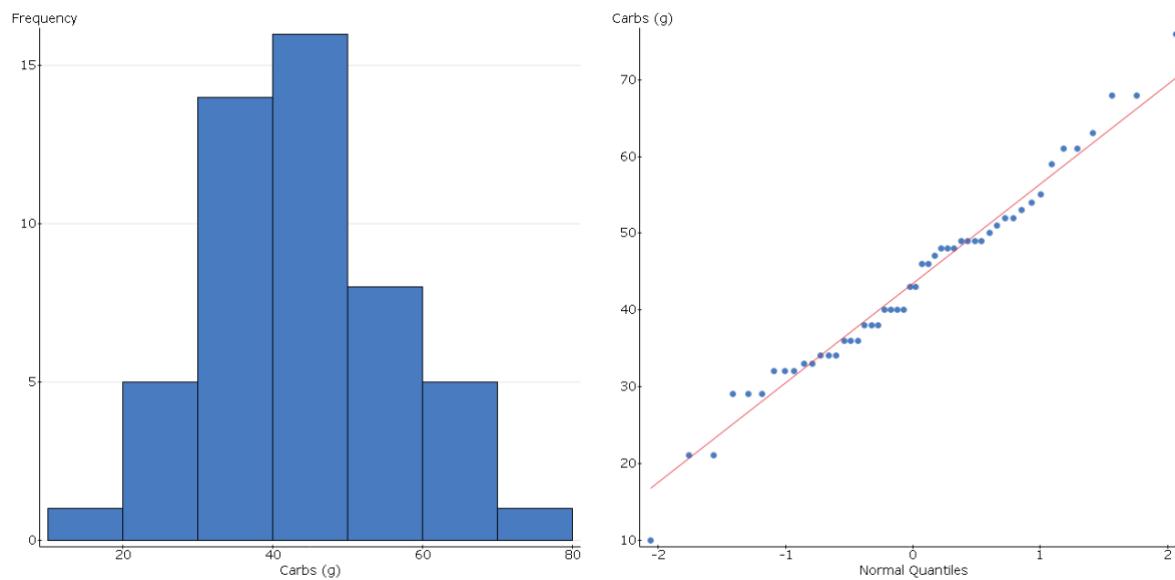
³ Here's how a QQ plot works. Organize your dataset smallest to largest. Then, turn your original data values into ordered pairs. Start with the smallest data value. On the x-axis, we plot a Standard Normal z-score for a dataset with the same number of data values that we have from the dataset we are investigating. On the y-axis goes the original data value. For instance, if you had a dataset with 99 values, we'd find the 1st percentile of the Standard Normal model, the 2nd, and all the way to the 99th percentile. If your variable is perfectly Normal, the dots will fall exactly on a straight, diagonal line. The further your variable is from Normal, the more the dots will depart from the straight line. Basically, the dots are lining up your data values with where they would be if they were Normal. This is not an exact science, so use your best judgment.

- You must check **both** a histogram and a QQ plot when making your determination for Normality. Sometimes, things are hidden in a histogram, like too much data piled up in the center or too-heavy tails. Always check both graphs.
- With small datasets, it is very difficult to determine if a Normal model is appropriate. Use common sense and always try to collect enough data to make a fair judgment.

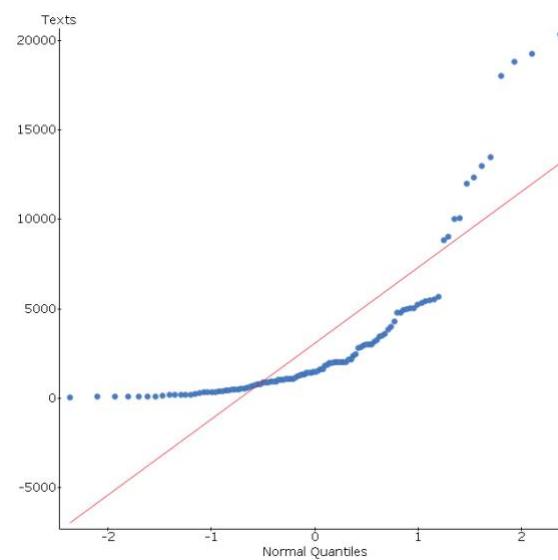
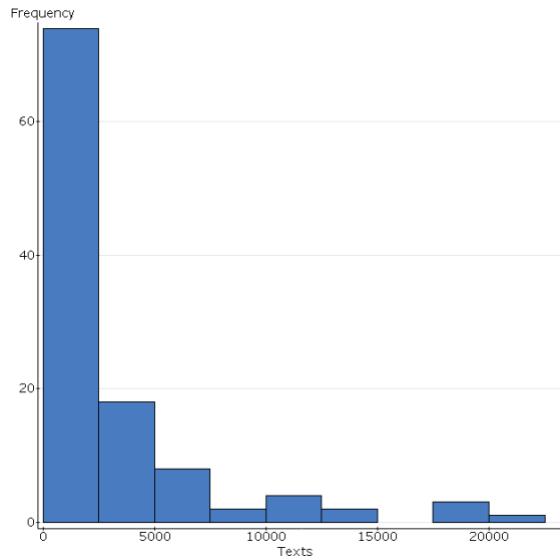
Example: Gas prices from 8 states were recorded over a period of months a few years ago. Based on the graphs, is a Normal model appropriate?



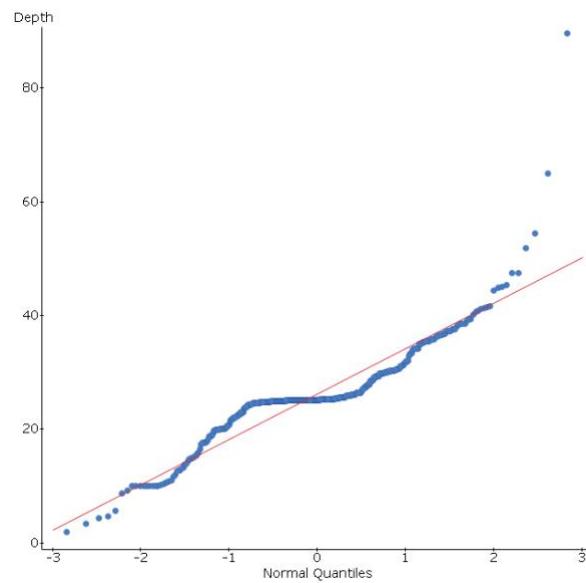
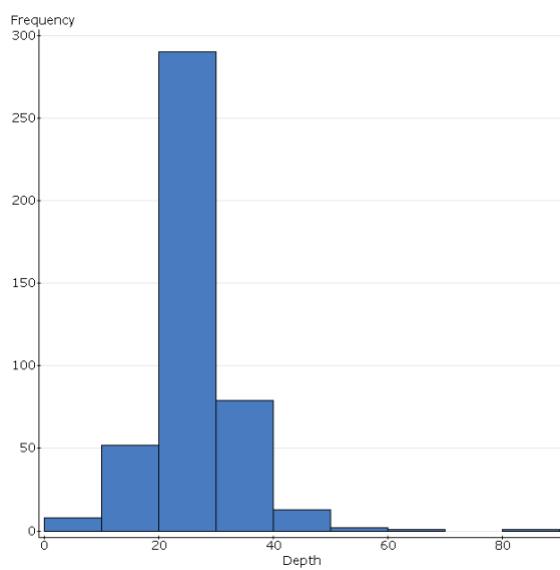
Example: Carbs were measured for a few dozen chicken sandwiches. Normal?



Example: Back in the Fall Semester, Cecil students reported the number of texts sent and received during the previous month. Normal?



Example: For the week surrounding the Honshu earthquake in Japan, every earthquake, no matter how small, was recorded by earthquake scientists. The variable was “*Depth*”. Normal?



Lesson 13: The Binomial Model

Blackboard Video 13-A

Example: Every time you drive to campus, you always try to get a parking spot in the bottom lot. Usually, you fail, but you keep trying. From long term experience, we know there is a 25% chance you get a spot down below. Your full schedule dictates that you come to campus five days each week.

- There are two possible outcomes each day, generically called _____ and _____. For this example:
- The **probability of success**, called _____, is the same for each trial. For this example:
- The **probability of failure**, called _____, is also the same for each trial. For this example:
- The trials are _____. In other words, whether or not you get a spot on Monday, you still have a 25% chance for Tuesday, Wednesday, etc....
- Situations like this are called _____.
- If we're interested in the **number of successes** out of n Bernoulli trials, we can use a _____.
- For this example, we're interested in the probability that we get a parking spot exactly two out of the five days. Exactly two successes means exactly three failures. ***It would be nice if the probability was (but it's not):***

- We need to take into account which two days we get that parking spot! There are many ways to get a spot exactly twice in one week.
- List out every possible scenario below.

M	T	W	R	F

- Each different arrangement in which we can have x successes in n trials is called a _____ . It is denoted as:
- The formula for nC_x is:
- $n! =$ $4! =$
- $1! =$ By Definition: $0! =$

- When we compute binomial probabilities by hand, we will need the value of nC_x . Let's practice:
 - a. Suppose we have 5 trials and are looking for the number of ways we can realize exactly 2 successes (parking lot example). Compute $5C_2$.
 - b. A group of 6 students would like to select 3 to go on a Taco Bell run. In how many ways can we select the 3 students who must get in the car and wait in the drive thru line for 45 minutes? Compute $6C_3$.

Blackboard Video 13-B

(Example Continued)

Every time you drive to campus, you always try to get a parking spot in the bottom lot. Usually, you fail, but you keep trying. From long term experience, we know there is a 25% chance you get a spot down below. Your full schedule dictates that you come to campus five days each week.

- **In general**, the probability of exactly x successes in n trials is:
- **For this example**, the probability of getting a spot down below exactly two days next week is:

- How many days each week should we expect to get a spot? That is the mean.
Intuitively, it is:

Formula for the mean of a binomial model:

The mean for the parking lot example:

- The standard deviation is not so intuitive. Trust us that for this model:

Formula for the standard deviation of a binomial model:

The standard deviation for the parking lot example:

- For the parking lot example, what is the probability that we get a spot on 4 of the 5 days?
- What is the probability that we don't get a spot at all during a week?
- Give the probability that we get a spot on at least one day during the week.

Blackboard Video 13-C

The Binomial Probability Model

Shorthand Notation: $\text{Binom}(n, p)$:

n = number of trials (fixed)

p = probability of success (constant)

$q = 1 - p$ = probability of failure (constant)

x = number of successes in n trials

The variable of interest is categorical, with two outcomes per independent trial.

The formula for x successes in n trials:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad 0 \leq x \leq n, \quad \text{where } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Mean: $\mu = np$

Standard Deviation: $\sigma = \sqrt{npq}$

Example: In September of 2013, just 14 percent of all Americans approved of the job Congress was doing. Maybe things are brighter here in Maryland. Pollsters take a sample of 20 residents and ask “Do you approve of the job Congress is doing?” We will count up the number who say “Yes”.

a. Verify that the conditions are met to use a binomial model.

b. Write out the binomial formula using the correct symbols.

c. What is the mean for this model? In other words, how many people do we expect to approve of Congress?

Example: A certain NHL superstar has a lifetime success rate of 64% on shootout attempts. Presuming attempts are independent and that he will receive 10 attempts this season, answer the following binomial model questions.

a. Confirm that a binomial model is appropriate by checking the conditions.

b. Give the binomial formula with the correct symbols.

c. How many shootout goals does he expect to score this season?

d. What is the standard deviation?

e. If he only scored 2 goals this season, is that unusually low? Why?

f. Determine the probability that he scores 9 goals.

g. Determine the probability that he scores 10 goals.

h. Determine the probability that he scores at least 9 goals.

Blackboard Video 13-D

- There is a StatCrunch tool to compute binomial probabilities (just like there is for Normal probabilities). Students are encouraged to use it, especially for problems with larger values of n and x (because the computations by hand will get tedious quickly).

Example: In September of 2013, just 14 percent of all Americans approved of the job Congress was doing. Maybe things are brighter here in Maryland. Pollsters take a sample of 20 residents and ask “Do you approve of the job Congress is doing?” We will count up the number who say “Yes”.

a. This is a binomial problem. Give the values of n and p and q .

b. Out of the 20 respondents, how many do we expect to support Congress?

c. Compute the value of the standard deviation.

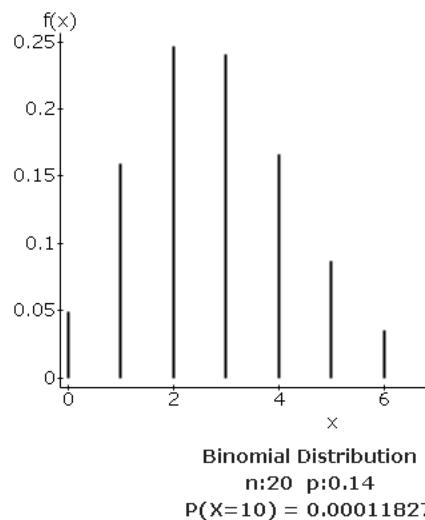
d. A pollster comes home from work one Saturday and is surprised because 10 of his 20 respondents said that they support Congress. Write out the binomial formula but use StatCrunch to compute the answer.

Binomial Formula:

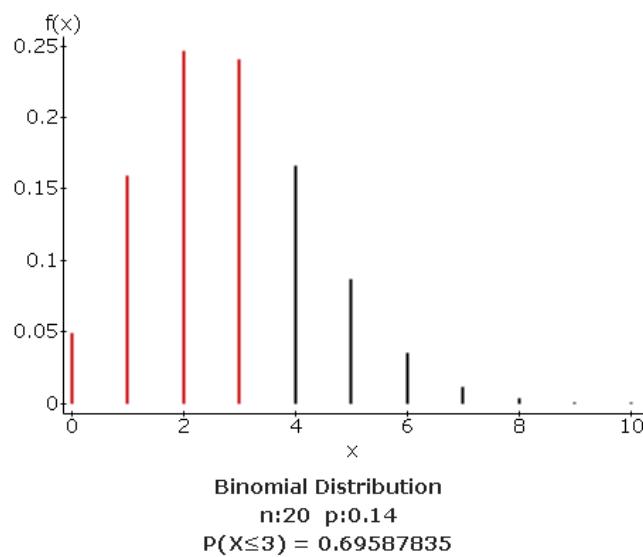
StatCrunch: Stat → Calculators → Binomial

Enter: n: 20 p: 0.14

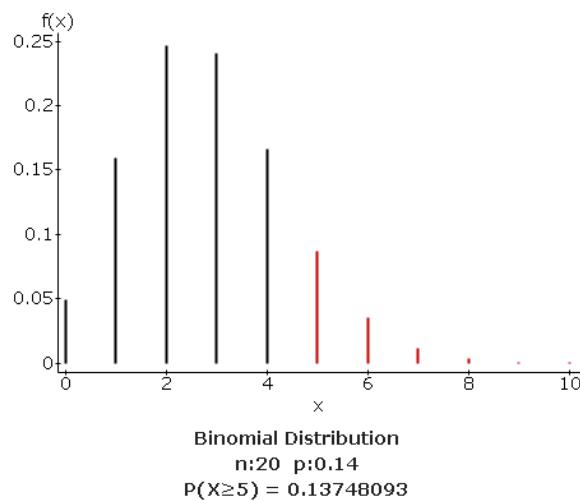
P(X = 10) Hit Compute



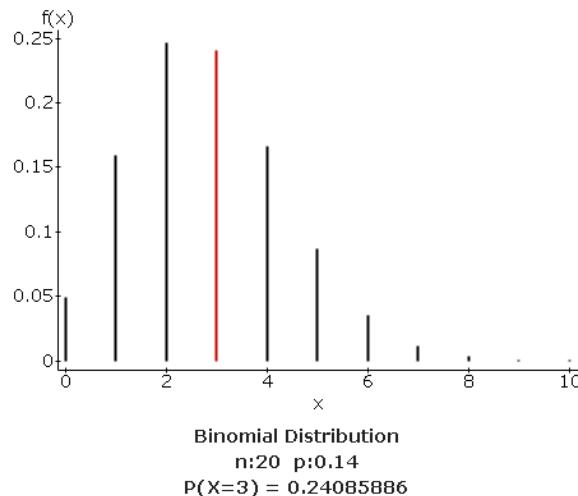
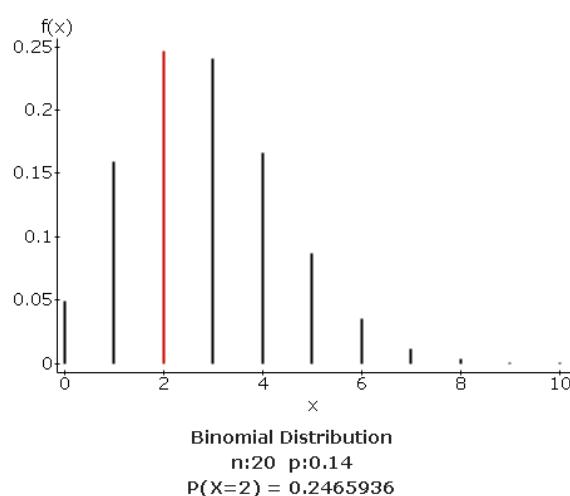
- Remember, we expect only 2.8 out of 20 to support Congress. To get 10 is exceptionally rare!
- e. What is the probability a pollster has at most 3 people supporting Congress? Write out the binomial notation, but use StatCrunch to obtain the answer.



- f. Give the probability that at least 5 people support Congress. Write the notation but use StatCrunch for the numerical answer.



- g. Determine the probability that a pollster finds 2 or 3 people that support Congress.

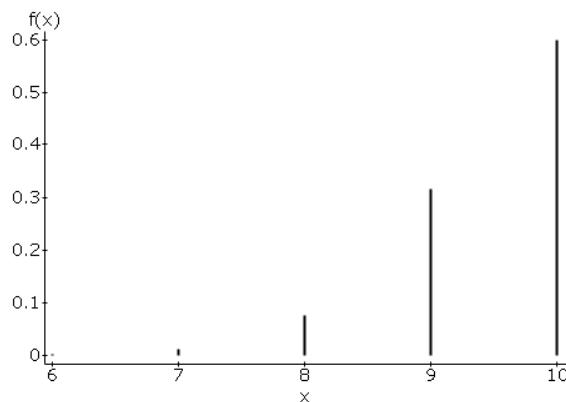


Blackboard Video 13-E

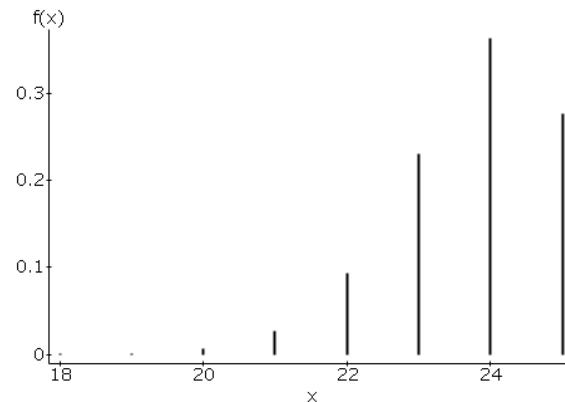
The Normal Approximation to the Binomial Model

- When the number of trials increases in a binomial probability model, something interesting happens. Check this out:

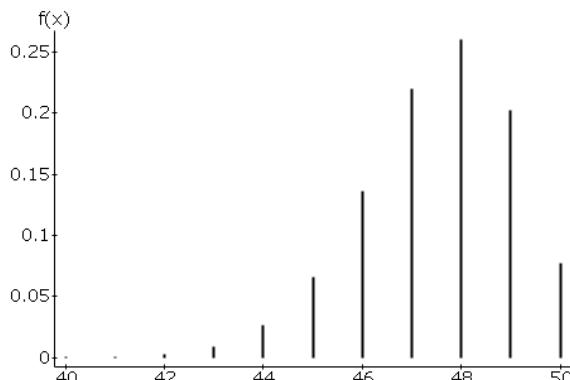
Example: A news blip claims that 95% of all gadgets returned to the store as being faulty actually still work just fine. Look at the binomial models for an increasing number of gadgets; we'd be counting up the number that still work just fine.



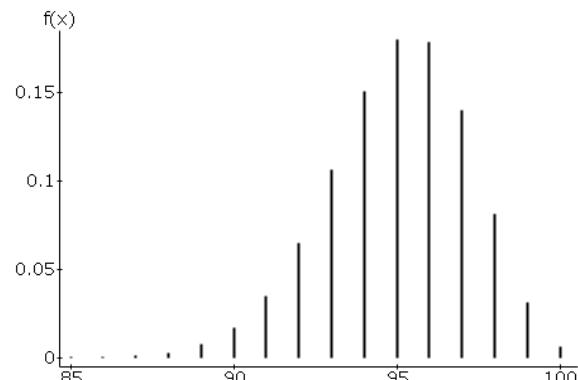
Binomial Distribution
n:10 p:0.95
 $P(X=\text{NaN}) = \text{NaN}$



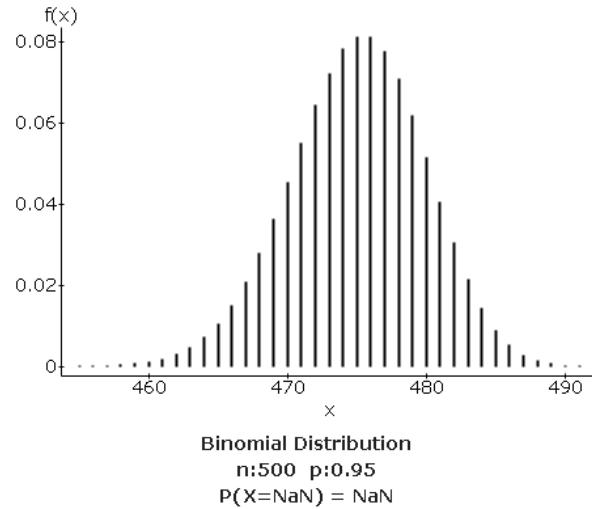
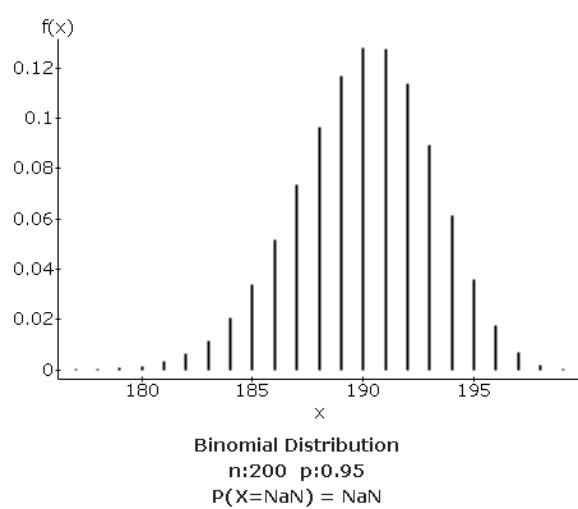
Binomial Distribution
n:25 p:0.95
 $P(X=\text{NaN}) = \text{NaN}$



Binomial Distribution
n:50 p:0.95
 $P(X=\text{NaN}) = \text{NaN}$



Binomial Distribution
n:100 p:0.95
 $P(X=\text{NaN}) = \text{NaN}$



- How will we know when it is OK to approximate a Binomial model with a Normal model? The reasons are a bit on the technical side for our purposes, but take our word for it.

The Normal model works fairly well as long as we expect to see at least 10 successes and 10 failures. We can call this the **Success / Failure Condition**. Here is the condition to check:

- When using a **Normal** model, we will need a mean and a standard deviation. Use these expressions:

Example: Suppose Wal-Mart expects 1000 returns during the month of January on gadgets, and they expect 95% of the returns to actually be in working order.

a. If we would like to compute probabilities for the number of returns that actually work properly, explain which model we would use and give the parameters and formula.

b. Why is it OK to approximate the binomial calculations with a Normal model?

c. Since a Normal model is OK to use, give the value of the mean and the value of the standard deviation.

- d. Give the probability that at least 965 out of the 1000 gadgets returned actually work properly (StatCrunch Normal calculator). Use the Normal approximation.
- e. Give the probability that at most 940 gadgets work properly. Use the Normal approximation.
- f. Give a range of values that describe an unusually low number of working gadgets. What about an unusually high number of working gadgets?

Math 127 Unit II Checklist

- I can describe the relationship between two quantitative variables using the proper terminology for form, direction, strength, and unusual features.
- I know how to make and interpret scatterplots, using StatCrunch.
- I know how to use StatCrunch to get the correlation coefficient, the regression equation, and R-Sq. I can interpret all these values in context.
- I understand not all relationships are linear, and that linear regression only works for linear relationships.
- I am aware of the effects of outliers, leverage points, and influence points.
- I can use a regression model to make predictions, and I am weary of extrapolation.
- I understand observational studies and regression models usually don't imply a cause-and-effect relationship.
- I understand that a straight line model should not be applied to curved data and I am aware of the techniques used to circumvent this issue.
- I understand the rules of probability and can apply them to different scenarios.
- I can draw and label a Normal model, and use StatCrunch to answer questions about a Normal model problem.
- I understand the Standard Normal model and can connect z -scores to a Normal model.
- I can set up the formula for a Binomial model problem, solve binomial problems using both the formula and StatCrunch.
- I can approximate a Binomial model using a Normal model when the conditions warrant. I understand how to check these conditions.

Lesson 14: Modeling the Sample Proportion

Blackboard Video 14-A

Example: At Cecil College, we are charged with the task of estimating the proportion of students who smoke cigarettes. Recently, the **nationwide** percentage is holding at 20%, but at Cecil College, we really wouldn't know until we collected some data.

We have a hunch, that at Cecil College, a greater percentage of our students are smokers when compared to the national figure. The hunch is based on our anecdotal evidence – it seems like we see a lot of people smoking, Cecil County is rural, and there are a lot of cigarette butts by the doorways and in the parking lot.

To really determine whether our hunch is valid, we'd need to collect data.

First, we have to define “**smoker**”, so we will say that if you have smoked at least one cigarette during the last 7 days, you are a smoker. Sorry if you don't like this definition, but it's what we're going with.

The **true proportion** of student smokers is _____ at

Cecil College. This value describes the population, which makes it a

_____. For this example, the population is

We decide that at Cecil College, we will take a sample of $n = 100$ students by randomly selecting names from Spring semester registration records and polling students when they pay for their classes. Once the data is collected, we can calculate the sample proportion, _____, from the student responses.

It turns out that our sample of 100 students had 30 smokers. That's 30%.

Think. What would happen to the 30% if we throw back those 100 students and randomly selected 100 new ones?

What if we did it again? And again. Each time, recording the value of the sample proportion of smokers?

This lesson investigates the theoretical distribution of all those sample proportions **if** we happened to take many, many repeated samples. Which we will never do. Because collecting data takes time and costs money.

Diagram the population and sample for the Cecil College smoking example:

Blackboard Video 14-B

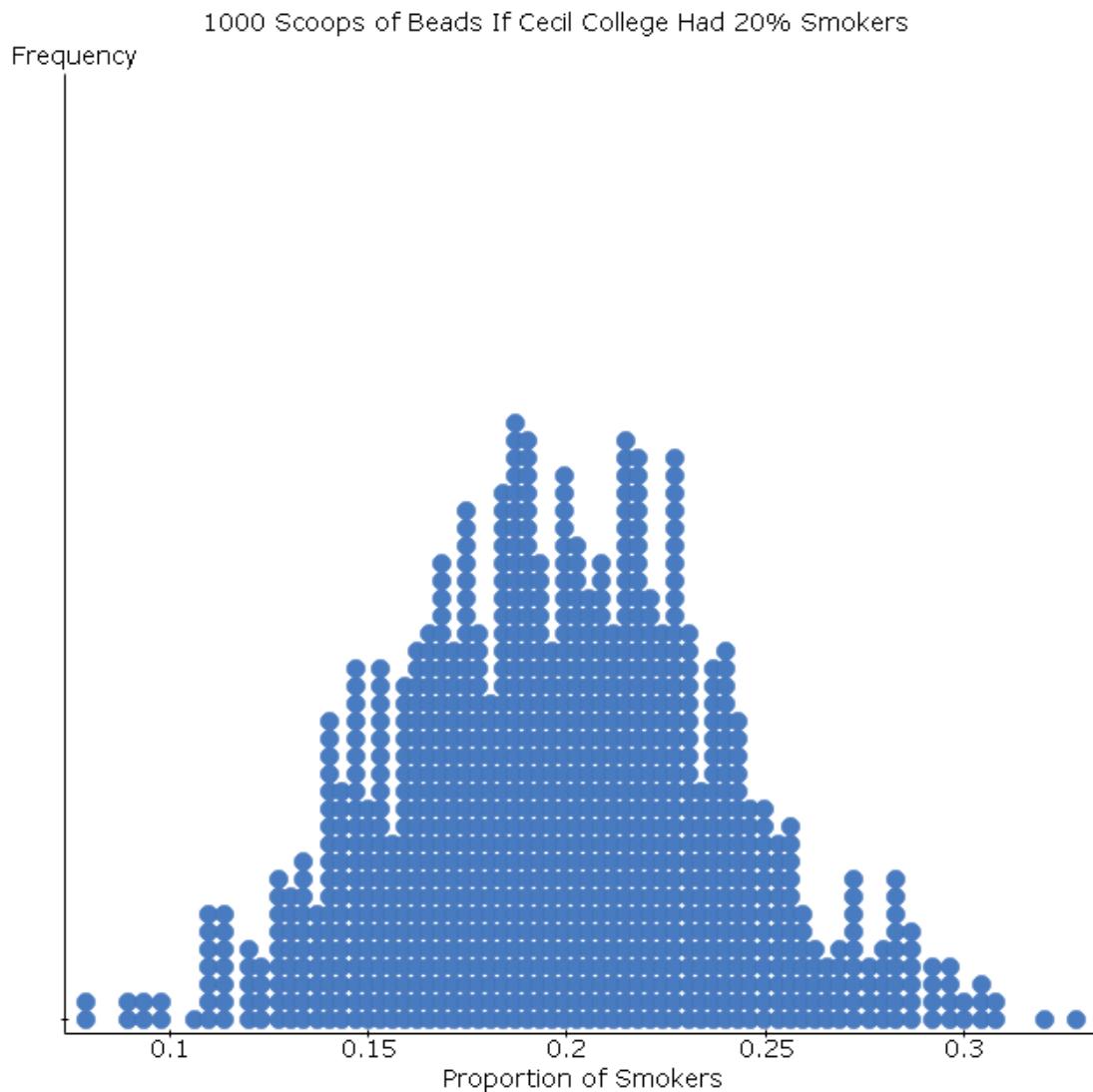
Demonstrate the idea of theoretically taking repeated samples using a manual simulation:

1. We have a box of over 2500 many-colored beads, with the green ones representing the smokers at Cecil College. We have no idea the true proportion of green beads in the box, just like at Cecil College, we have no idea of the true proportion of smokers for the whole college. We might guess at Cecil College it's 20% smokers, because that's the national proportion.
2. The paddle has 100 slots, so we can quickly take a simple-random sample of 100 students, count up the green beads representing the smokers, and determine the proportion in the sample who were "smokers".
3. We will take 10 samples of size $n = 100$ and record the proportion of smokers below:

Sampl e 1	Sampl e 2	Sampl e 3	Sampl e 4	Sampl e 5	Sampl e 6	Sampl e 7	Sampl e 8	Sampl e 9	Sampl e 10

4. What we are investigating is the behavior of the sample proportion when we take repeated random samples. Remember, we are investigating the hunch that at Cecil College, more than 20% of all students smoke. Since we will only take one sample in practice, we need to understand the variation and behavior of the sample proportion, so that we can make the most out of the one actual sample we will collect.
5. Where should all the sample proportion values in the boxes center around? In other words, what should the mean be?

6. At this point (we will do simulations in class), take our word for it that if you kept scooping paddles of beads and keeping track of all the sample proportions, a dot plot or histogram of the values would start to look pretty darn **unimodal** and **symmetric**:



- What model would you use for this data? _____
- When working with this model, what are the two model parameters that we need to know?
_____ and _____

7. So, under certain conditions, the sample proportion will be well-modeled by a Normal model.

The mean of the sampling distribution model for the sample proportion is:

The standard deviation of the sampling distribution model for the sample proportion is:

8. Now, at Cecil College, we do not know the proportion of student smokers, so we will assume that it's 20%. We might change our mind if our one sample of data is compelling enough, but presuming it's 20%, what is the correct model we can use to model the sample proportion?

Blackboard Video 14-C

Example: At Cecil College, we assume 20% of students smoke, but have a hunch that more than 20% of students smoke. We'd like to model the sample proportion with a Normal model. Determine the mean and standard deviation of that model.

Conditions We Must Always Check to use a Normal Model for a Sample Proportion

1. **Categorical Variable Condition:** The variable of interest must be categorical.
2. **Randomization Condition:** Attempt for a _____.
Realistically, this is difficult, so be confident the sample is not _____.
3. **Sample Size Condition:** The sample size shouldn't be more than 10% of the population size. We can check this in two equivalent ways:
4. **Success / Failure Condition:** For our Normal model to be accurate, we need to expect to have at least 10 successes and 10 failures (CC smoking example, we need to expect to sample at least 10 smokers, 10 nonsmokers). We can check this using:

Why Do We Need To Check Conditions to Use a Normal Model for the Sample Proportion?

1. We are modeling a sample proportion, which requires that our data is categorical. Further on in Math 127, we will model a sample mean, and the techniques will be slightly different. Always think about the type of variable you are working with because there are certain methods for certain data types!
2. Ideally, the data will be generated by taking a true simple random sample. But the real world gets in the way of data collection – if your data are biased, perhaps for the Cecil College smoking example we sampled only night students, or we smell-tested student's clothes, or who knows what, we won't have faith in our results. Plan out a sound data collection strategy every time.
3. The sample size must not get too large when compared to the population size. If it does, we might run into problems with data values being independent of each other. The statistical methods we use to perform inference will always require that every data value is independent of each other. This is how we check for that.
4. We need to expect at least 10 successes and 10 failures for our categorical variable. Think of that dotplot in Video Lesson 14-B where you visualized a Normal model from the simulated data. For technical reasons, if we don't expect at least 10 and 10, repeated samples won't have that unimodal and symmetric shape. In other words, a Normal model will be a bad fit, and we're using the Normal model to make decisions about our population of interest.

Cecil College Smoking Example

Verify the conditions to use a Normal model for the sample proportion:

For the **Cecil College** smoking example, draw the sampling distribution model for the sample proportion. Remember, we are assuming 20% are smokers, until the data convinces us otherwise. Label with the empirical / 68-95-99.7% rule.

- a. If at Cecil College, our one and only sample had 30% smokers, what should we conclude about Cecil College students and their smoking habits?

- b.** Convert the sample proportion of 30% to a z -score. Comment.
- c.** Use the Normal calculator on StatCrunch to determine the probability of collecting a sample of size 100 with 30% or even more than 30% smokers. Comment.

Blackboard Video 14-D

Example: One polling group indicates that 1.2% of adults in America can be classified as compulsive gamblers. We will investigate if the proportion has changed here in Maryland, anticipating **an increase** due to the proliferation of lotteries and legalized casino gambling nationwide during the past decade.

Researchers collect a random sample of 1200 Marylanders.

- a. Determine the correct model for the sample proportion. Check conditions. Draw and label the model.

- b.** It turns out that 15 adults were classified as compulsive gamblers in the study. Determine the value of the sample proportion.
- c.** Convert the sample proportion to a z -score and comment on its value.
- d.** Using the Normal model calculator on StatCrunch, determine and comment on the probability of our collecting a sample like we did, or one even more unusual, if our Normal model based on 1.2% problem gamblers is a good model for Maryland.

Blackboard Video 14-E

Example: In 2013, *Mother Jones* magazine reported that 45% of all homes in the US contained guns. We'd like to run an investigation in Cecil County to see if there's evidence that we differ substantially.

We will design a study to unbiasedly sample 400 homes in the county, and we will determine whether or not the homes selected have guns.

- a. Check all conditions so that we can proceed with modeling the sample proportion.
 - b. Determine the mean and standard deviation of the Normal model we will use for this problem.

c. Draw and label the Normal model.

d. Now imagine we are going to collect our data. What values of the sample proportion will lead you to conclude that Cecil County differs from the rest of the nation with respect to gun ownership? What values would lead us to opposite conclusions?

Blackboard Video 14-F

Wrapping Up Lesson 14 – A Worthy Review

- We are moving onto statistical inference, which entails creating _____ and running _____.
- In Lessons 15 – 17 we will be analyzing _____ variables.
When working with this type of variable, we will be using sample _____.
- To perform statistical inference, there will always be a theoretical model in the background on which all the methods are based.
- The sample proportion, when certain conditions are met, will closely follow a _____.
- The mean of this theoretical model is: _____
- The standard deviation of this theoretical model is: _____
- We will have to meet certain conditions before proceeding, so we will be sure to check:

- We will typically determine the theoretical model for our statistic of interest based on certain assumptions about our population of interest. This model may or may not be correct, based on the quality of our assumptions.
- It is our job to then collect data and weigh the data against the model. The data might convince us that our theoretical model is faulty, and this will lead us to change our minds about the population.
- Other times, the data won't convince us that our theoretical model is faulty. In these cases, we may continue to just go along with our assumptions that we made before collecting data.
- Recall the Cecil College smoking example. We began our investigation by assuming that 20% of all Cecil College students were smokers. We cooked up the correct Normal model based on this assumption.
- Once we collected our data, we determined just how unusual it would be to obtain our data, if the Normal model based on $p = 0.20$ is reasonable for Cecil College.
- Then we made a conclusion about the entire population, just based on our small sample.
- When the collected sample proportion falls far out in the tails of the Normal model, then the data have convinced us to change our minds about the whole population of Cecil College students.
- “Far in the tails” really depends on the exact Normal model we are using. We have to recall our abilities to work with z -scores and to shade under Normal models to obtain probabilities.
- Probability comes into play, and essentially we will be turning all of these investigations into probability problems – “What is the probability that I just drew my sample, or one even stranger, if this model is good?”
- The rest of Math 127 is based on the main ideas of Lesson 14, so be sure to get questions answered from your instructor as we continue on in the course.

Lesson 15: Confidence Intervals for Proportions

Blackboard Video 15-A

Example: Recall that in 2013, *Mother Jones* magazine reported that 45% of all homes in the US contained guns. We'd like to run an investigation in Cecil County to see if there's evidence that we differ substantially.

We will design a study to unbiasedly sample 400 homes in the county, and we will determine whether or not the homes selected have guns.

- a. In Lesson 14, we looked at the theoretical Normal model of the sample proportion, if we *imagine* taking many, many repeated samples. That model was based on one crucial assumption:

- b. Now, in Cecil County, we would like to estimate the proportion of homes with guns. We will take a sample of 400 homes. It turns out that a whopping 300 homes have guns (we made this up, but just go with it). If you had to give your *best guess* for the gun ownership rate in the **whole county**, what would it be?

- c. We'd like to do better than a single best guess. In Lesson 15, rather than give a single best guess, we'd like to give a range of _____ values. This is called a _____. We will use the ideas from Lesson 14 to build this interval.

- d. We know that in Cecil County, the proportion of homes with guns will also follow a Normal model (if we took repeated samples), but $\hat{p} = \frac{300}{400} = 0.75$ is likely not the center of that Normal model. In Cecil County, we are actually working with a Normal model with an _____ mean. This is the first time we've been in this situation, and if we took another sample, surely we wouldn't get 300 gun owners again.

e. Where should we center our Normal model for the gun ownership rate in Cecil County?

f. When working with Normal models, we also need a measure of spread. Since we do not know the true proportion of gun ownership in Cecil County, we cannot obtain the **standard deviation** for the Normal model:

Since p is unknown, _____

Our solution is to use the collected data, the sample proportion, to estimate a measure of spread for our Normal model:

We know $\hat{p} = \frac{300}{400} = 0.75$ from our data, so use this value to estimate the **standard error** in place of the standard deviation:

g. **Keep focused.** We are trying to come up with a range of plausible values for the rate of gun ownership in Cecil County. We have a best guess at 75%. We now have an estimate of the spread, which is the standard error of 2.17%. We also know that for sample proportions, a Normal model applies.

Let's use some Normal model theory to build our confidence interval.

- h. The **empirical rule** for Normal models tells us that:

_____ of samples will produce a sample proportion within **one** standard error of the true proportion of gun owners in Cecil County.

_____ of samples will produce a sample proportion within **two** standard errors of the true proportion of gun owners in Cecil County.

_____ of samples will produce a sample proportion within **three** standard error of the true proportion of gun owners in Cecil County.

- i. In practice, 95% confidence intervals are very common. Build a 95% confidence interval for the true proportion of gun owners in Cecil County.

- j. Finally, we need to interpret our interval with a sentence in context:

- k. Does it look like we have compelling statistical evidence that the gun ownership rate in Cecil County differs from the national figure of 45%? Why?

Blackboard Video 15-B

The Official Steps to Calculate a Confidence Interval for a Proportion

1. Check the conditions. They must be met to proceed.
 - A. Variable is _____ because we're working with proportions.
 - B. The sample is ideally _____ but at the very least, _____.
 - C. The size of the sample is less than _____ of the size of the population.
 - D. We have at least _____ successes and _____ failures in our collected data.
2. Based on the sample data, determine the value of \hat{p} .
3. Using the Standard Normal model, determine the value of z . Details on the next page. The number z is called a **critical value** and is based on the confidence level for your interval.
4. Calculate the confidence interval using the following formula:
5. Interpret your interval with a sentence in context.

Finding the Critical Value z

- Use the StatCrunch Normal model calculator, but because these values are commonly used, write them down somewhere you can find them quickly.
 - In practice, 95% and 99% confidence intervals are the most common. Less common in practice, we will also determine the critical values for 90% and 98%. If you're ever asked to find a critical value for a weird confidence level, like 93%, you'd follow the same steps.
- a. Determine the critical value z for a 95% confidence interval for a proportion. Use the StatCrunch tool.

b. Determine the critical value z for a 99% confidence interval for a proportion.

c. Determine the critical value z for a 90% confidence interval for a proportion.

d. Determine the critical value z for a 98% confidence interval for a proportion.

Blackboard Video 15-C

Two Examples

1. Recall from Lesson 14 we looked at the percentage of the population that are classified as “compulsive gamblers”. The last reported value was 1.2%, but here in Maryland, we had taken a random sample of 1200 adults and diagnosed 15 as compulsive gamblers.

Create a 99% confidence interval for the true proportion of Maryland adults who are compulsive gamblers. Check conditions first.

- 2.** Data was collected on campus to ascertain whether or not students are using the physical education complex (for things other than their regularly scheduled classes). Staff members visited an assortment of classes and polled the students. Altogether, it turns out that 46 students use the facility and 230 students don't use it.

Calculate a 98% confidence interval for the proportion of all Cecil students using the gym. Interpret. Does it seem the college is meeting its goal of at least 25% of students using the facility?

Blackboard Video 15-D

Facts About Confidence Intervals

1. The **parameter** we are estimating, in Lesson 15 it's a population proportion, doesn't vary, and even after we collect our sample, we still won't know its true value. We could say the parameter is _____ and _____.

Example: In Video 15-C, we estimated the proportion of Maryland adults who are compulsive gamblers. Even after we collected our data from 1200 adults, we still don't know the true proportion for all adults in the entire state! We'd have to collect data from millions of people to obtain that value!

2. If we were to take another sample, we'd get different data, which would result in a different confidence interval. Our confidence intervals are _____, based on the random sample we happened to collect during the data gathering process.

Example: When we created our 99% confidence interval for the true proportion of compulsive gamblers in Maryland, we had 15 out of 1200 diagnosed. The interval was (0.42% to 2.08%) for the whole state.

What if our random sample contained 25 problem gamblers? For practice, check at home that the 99% confidence interval is (1.021% to 3.145%). Different data, different interval. In practice we take just one sample and use it to the best of our statistical abilities.

3. The word confidence refers to something you might not expect it to be. It is not a probability. Rather, it is the proportion of intervals that would actually contain the parameter we are estimating, ***if we took many, many repeated samples***, which we won't.

Example: Suppose in Maryland, that actually 2% of all adults are compulsive gamblers.

Then, when we take a sample to create a confidence interval, most of our intervals will contain that target 2%, but a few of them won't.

If we decide to create a 99% confidence interval, then in the long run, _____ of the intervals will hit the target. 1% will miss.

Unfortunately, we won't know if our particular interval hit or missed, so the best we can say is "I'm 99% confident that my interval captures the parameter I'm estimating".

4. The more confident you need to be (of containing the parameter you're estimating), the _____ we must purposefully make the interval. When we change the confidence level, the _____ will change in our formula:

$$\hat{p} \pm z \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: In our Cecil County gun ownership example, we had 300 homes with guns from our sample of 400. The 95% confidence interval was (70.757% to 79.243%). What if, instead we required 99% confidence? Determine the interval and note that it is wider.

5. The quantity to the right of the plus / minus sign is called the _____. We like it to be _____ because then our intervals are skinnier and more useful. Give the formula for margin of error for a confidence interval for one-proportion:

You can see from the formula, that margin of error is based on three quantities:

Margin of error depends on _____, _____, and _____.

- We cannot change \hat{p} because that value came from our collected data.
 - We can change the sample size, n , by collecting more or less data. Be advised that collecting data takes _____ and costs _____, (once you leave the classroom setting).
 - If you increase the sample size, what happens to margin of error? Look to the formula:
 - We already learned that we can also change the confidence level. If we need to be more confident in our interval, the margin of error will get _____ and if we can live with less confidence, the margin of error will get _____. Again, look to the formula:

- 6.** We can make conclusions about the population based on our sample and the confidence interval we calculated from the sample. Check to see if the population parameter of interest falls **inside** or **outside** of your interval.

Example: In the Lesson 15 opening example, Mother Jones magazine reported that 45% of all households in the US contain guns. Based on our 95% confidence interval from 400 Cecil homes, do we have evidence that Cecil differs from the rest of the country? The interval was (70.575% to 79.243%).

Example: In the gambling example, it was stated that 1.2% of all Americans are compulsive gamblers. The 99% confidence interval for Maryland residents was (0.42% to 2.08%). What can we conclude about Marylanders with respect to the rest of the nation?

- 7.** Always make your concluding remark a statement about the entire _____. We are doing statistical inference, always inferring back to the larger group.
- 8.** If you're making a conclusion about a population using a confidence interval, do not fiddle with the confidence level to be able to make the conclusion you want to make. Pick your confidence level _____, collect your data _____, and finally make your conclusion.

Blackboard Video 15-E

Determining Sample Size When Working With Proportions

- Before engaging in a statistical study to investigate a population, care should be taken to collect a sample size that will meet your needs. A few things to consider:
 1. What _____ can you live with? If your interval is too wide, it will be difficult to make useful decisions with it.
 2. What _____ do you require? The main two choices are 95% and 99%. If there are severe consequences when creating intervals that do not capture the parameter we are estimating, then be more confident.
- The sample size formula comes from algebraically manipulating the margin of error formula:

- The values of z and ME come from your requirements in **1** and **2** above. What about \hat{p} ? If you have an educated-guess for \hat{p} , then use that value. An educated-guess is not a random guess!

If you have no idea what \hat{p} might be, default to $\hat{p} = \underline{\hspace{2cm}}$. This is the most conservative value we can put and will lead to the largest sample size for whatever your z and ME are.

Example: At Cecil College, we'd like to estimate the proportion of students who would vote Democrat. A quick Internet search reveals that Maryland is 56% Democrat, and no immediate information was available for the county. We feel a 4% margin of error will work for our purposes and are happy with 95% confidence.

- a. Determine the required sample size if we are willing to assume 56% Democrat is a good starting point for Cecil College.
 - b. Now determine the required sample size if we are not willing to assume 56% is a good starting point for Cecil College.

Example: In Cecil County, we'd like to estimate the proportion of homeowners that are currently behind in their payments. County officials would like to be within 3 percentage points and 99% confident with their interval estimation.

- a. How large a sample size ought they take? We learn from a Zillow report that currently, 21.5% of all Americans are late on the payments.

- b. What if officials want to reduce margin of error to 1%? Give the sample size.

Blackboard Video 16-A

Running a Hypothesis Test for One Proportion

Example: A few years ago, an ABC News poll stated that 44% of Americans surveyed said that they drove an SUV, minivan, or some kind of truck. Here at the college, we notice that almost everything in the parking lot seems to be something bigger than a car.

We will run a hypothesis test to see if more than 44% of Cecil College drivers prefer larger vehicles.

Students roam the parking lot on various days using good sampling techniques: 100 large vehicles, 90 small vehicles. The data can be summarized as

- We need a **temporary working model**, so we will _____ that at Cecil College, 44% of our drivers prefer large vehicles. This assumption may be correct or incorrect, but we have to start somewhere.
- We will _____ whether or not, based on our sample, that there is compelling evidence that **more than** 44% of our drivers prefer large vehicles.
- The first step is to write down the **null** and the **alternative** hypotheses:

Null Hypothesis: _____

Alternative Hypothesis: _____

- **Ask yourself this question:** “What would convince you that more than 44% of **all** Cecil College drivers prefer large vehicles?”
 - For one, our sample proportion would certainly have to be somewhat _____ than 44%. **But how much larger?**
 - Our temporary working model, the one based on our assumption that _____ can help us make the call.

- The standard deviation of the sampling model is the key. We calculate it based on our initial assumption that $p = 0.44$:
- Using the assumed model, centered at $p = 0.44$, with a standard deviation of $SD(\hat{p}) = 0.036$, we can ask ourselves **just how unlikely** it would be to actually draw our sample proportion of $\hat{p} = \frac{100}{190} = 0.526$.
- Typically, when running a hypothesis test, the sample proportion gets standardized, so we will do that now. Convert the sample proportion to a z -score:
- As a friendly reminder, interpret the z -score with a sentence in context:

- When running hypothesis tests, we will typically base our decisions on probability rather than the standardized value. Determine the probability the we collected our sample, with $\hat{p} = \frac{100}{190} = 0.526$ or 52.6% large vehicles, or a sample even more extreme, if our initial assumption that 44% of all vehicles are large, is true. (Read the preceding sentence until you understand what it's saying. It's quite important.)

- Now it's judgment call time. One of two possibilities is true.
 - The null hypothesis is true. Our initial assumption that 44% of Cecil College drivers prefer large vehicles is correct. AND we just drew a sample that would only occur 0.84% of the time (about 8 in 1000 or about 1 in 125 times).
 - Because the data we collected has such a low probability of occurring, if the null is true, then in all likelihood, the null is not true. Rather, the alternative hypothesis is likely true. The data have convinced us to change our minds. The $\hat{p} = 52.6\%$ was compellingly higher than our initial 44% assumption, so we conclude that at Cecil College, more than 44% of all drivers prefer large vehicles.
- We are convinced to select option _____, and feel comfortable about that decision.

Important Considerations About This Hypothesis Test

1. We are not 100% certain that more than 44% of all Cecil College drivers prefer large vehicles. In fact, with our decision to reject the null and go with the alternative hypothesis, there is a _____ chance we made a mistake. If that's the case, then we've come to the wrong conclusion.

Because when running a hypothesis test we are using _____ information about the population, namely the _____ data, there is always a chance of making a mistake. The chance of making a mistake, though, with good data collection and an ample sample size, will usually be quite small.

2. We've concluded that more than 44% of Cecil College drivers prefer large vehicles, but we still do not know exactly what the percentage is. We'd have to take a _____ of every vehicle to obtain the true percentage. Be careful and don't conclude that at Cecil, 52.6% of all drivers prefer large vehicles.
3. If we'd like to give a range of plausible values for the true proportion of those preferring large vehicles, we can support our hypothesis test with a confidence interval:

Blackboard Video 16-B

The General Steps to Run Any Hypothesis Test

In Math 127, we will run many types of hypothesis tests, depending on if the variable is categorical or quantitative, and depending on if we are investigating one or two samples. You can always follow this model to make sure you hit all the steps, though the details might change depending on the situation.

1. Read and understand the context of the problem. Determine if the variable is categorical or quantitative and if you're working with proportions or means. Determine if you are dealing with one or two samples.
2. Write down the hypotheses.
3. Check the appropriate conditions before proceeding.
4. Based on the collected sample data, convert your sample proportion(s) or sample mean(s) to a test statistic.
5. Determine the probability that you got your test statistic, or one even more unusual, if the null hypothesis is correct. This is called the _____.
6. If the P-Value is small, then you'll reject the null hypothesis in favor of the alternative.
7. Write a concluding remark in the context of the problem.

Official Steps to Run a One-Proportion Z-Test

1. Write the hypotheses.

Null Hypothesis:

Alternative Hypothesis:

2. Check conditions. If they are met, then a Normal model will be appropriate to use for the test.

The variable is _____.

The sample is _____ or at least _____.

The sample size n is less than _____ of the population size N .

We expect at least _____ successes and _____ failures. You can check:

3. Run the test. **First**, convert your data to the test statistic:

Second, get the P-Value by shading under the Standard Normal model in the H_A direction:

4. Make your decision. If the P-Value is “small”, reject the null in favor of the alternative. If the P-Value is “large”, fail to reject the null hypothesis.
5. Write a concluding remark, based on your decision, in the context of the problem.

P-Values and the Logic of Hypothesis Testing

- Here is the fundamental question behind every hypothesis test: “Are my data surprising if the null hypothesis is true?” When your data are surprising enough, we reject the null hypothesis in favor of the alternative.
- The probability of seeing data like ours (from our sample) if the null hypothesis is true is the P-Value. The _____ the P-Value, the more likely it is that the alternative hypothesis is the correct hypothesis.
- How small? P-Values are probabilities, so they range in value from _____ up to _____.
- In almost all cases, a P-Value of _____ or above is not showing much evidence in favor of the alternative hypothesis. The collected data are plausible under the assumption that the null hypothesis is true. We would _____ the null hypothesis.
- In almost all cases, a P-Value of _____ or below is showing strong evidence in favor of the alternative hypothesis. The collected data are not plausible under the assumption that the null hypothesis is true. We would _____ the null hypothesis in favor of the alternative.

P-Value Diagram

- **IMPORTANT:** If you do not reject the null hypothesis, you have not proven the null hypothesis to be true.

Blackboard Video 16-C

Example: Education-Portal.com has the following interesting figures (true?):

- Cheating most often occurs in math and science classes.
- 75 to 98 percent of college students surveyed each year admit to cheating at some time in their college careers.
- The college students most likely to cheat are engineering and business majors.

Let test for evidence that more than 75% of the students at a given school cheat. We randomly, anonymously, and confidentially collect data from 444 students. It turns out that 360 admit to some form of cheating during their college career. Run the test, showing all steps.

1. Read and understand the context of the problem. Determine if the variable is categorical or quantitative and if you're working with proportions or means. Determine if you are dealing with one or two samples.
2. Write down the hypotheses.
3. Check the appropriate conditions before proceeding.

4. Based on the collected sample data, convert your sample proportion to a test statistic.

5. Determine the P-Value.

6. If the P-Value is small, then you'll reject the null hypothesis in favor of the alternative.

7. Write a concluding remark in the context of the problem.

Blackboard Video 16-D

Example: One study reports that, historically, 34% of high school students have perfect attendance for the entire year. Suppose at Rising Sun high school last year, the principal became concerned when only 320 out of 1000 students had perfect attendance. Should the principal be concerned or might this just be year-to-year fluctuation?

- a. Run the one-sample z -test for a proportion to test this claim. Show all steps.

- b.** Interpret the P-Value with a sentence in context.

Blackboard Video 16-E

Making Errors when Hypothesis Testing⁴

- Whenever you run a hypothesis test, you can make a mistake. That's because you're using a relatively small sample to make a decision about a relatively large population.
- Sometimes, the data you collect, just by taking a random or unbiased sample, won't jive with the population. In other words, you happened to be unlucky when collecting your data.
- The problem is, we won't know if we got unlucky. We just have to be aware of the possibility.

Example: Suppose we work for Consumer Reports Magazine, and next month, we plan on recommending the Continental DWS tires as a good buy unless there is evidence that they don't last as long as the company claims (50,000 miles under normal driving conditions).

We buy 20 tires in total from various online outlets and run them out on our testing equipment.

In reality, four scenarios could happen.

- a. Draw a diagram of the population and the sample. Comment.

⁴ This is a problem about means, not proportions. The ideas are what matter. Means are discussed in Lesson 20.

- b. Write down the hypotheses that Consumer Reports are testing.

- c. Scenario 1. The null hypothesis is true and nothing in our sample refutes it.

Continental has made a correct claim. The tires last, on average, at least 50,000 miles. The data we collect at Consumer Reports is consistent with the claim. For example, the sample mean from our $n = 20$ tires turned out to be $\bar{y} = 51,558$. We recommend the tires in the magazine. Everyone is happy.

Which hypothesis is true? _____.

Based on the data, what is our decision? _____.

This is a _____.

- d. Scenario 2. The null hypothesis is true, but we collected an unlucky sample.

Continental has made a correct claim, but the sample of $n = 20$ tires had an unusually low sample mean. For whatever reason, we obtained $\bar{y} = 45,110$. This was so below 50,000, that we don't recommend the tires.

The problem is, we made a mistake based on data not consistent with the claim. Once in awhile, you get a weird sample, and this is one of those times.

Which hypothesis is true? _____.

Based on the data, what is our decision? _____.

This is an _____.

We rejected the null hypothesis, even though it was true. Continental will not be happy with us – They should have gotten a good recommendation but because of our unlucky random sample, they did not. This unfortunate result is a consequence of making decisions about a population based on incomplete information (i.e. a sample).

- e. Scenario 3. The alternative hypothesis is true and the sample is consistent with the alternative hypothesis.

Continental's claim that the tires last at least 50,000 miles is simply not true. Our sample of $n = 20$ tires indicates the same thing. For example, the data we collect has $\bar{y} = 45,653$. Consumer Reports magazine doesn't recommend the tires.

Which hypothesis is true? _____.

Based on the data, what is our decision? _____.

This is a _____.

Maybe Continental is unhappy, but they need to check their tires or adjust their claim.

- f. Scenario 4. The alternative hypothesis is true, but we collected an unlucky sample.

Again, Continental's claim about the tires is untrue, as the tires don't last the reported 50,000 miles on average. This time, however, our sample of $n = 20$ tires had an unusually high sample mean of $\bar{y} = 50,103$. With a sample mean exceeding 50,000, there is no way we can reject Continental's claim. We recommend the tires in the magazine.

Which hypothesis is true? _____.

Based on the data, what is our decision? _____.

This is an _____.

Now our reputation is on the line, and in no time, angry consumers who bought the tires on Consumer Report's recommendation will come looking for answers!

- g. Summarize, in general, the four possibilities when running a hypothesis test.

Blackboard Video 16-F

Example: Recall our college cheating example where we tested $H_0: p = 0.75$ versus $H_A: p > 0.75$ to determine if more than 75% of all students cheat at a certain college.

We obtained a P-Value of 0.0015 and decided to reject the null in favor of the alternative.

What kind of error could we have made? What are the chances we did? Explain.

Example: Recall our high school attendance example where we tested $H_0: p = 0.34$ versus $H_A: p < 0.34$ to determine if the perfect attendance rate has declined below 34%.

We obtained a P-Value of 0.0901 and failed to reject the null.

What kind of error could we have made? Explain.

Example: An allergist in Indianapolis speculated that perhaps the reason why allergies and asthma have been on the rise over the past 100 or so years is because people are no longer exposed to farms like they were in the past. He hypothesizes that exposure to farms, particularly cows, influences the bacteria in our guts, which “stimulates” the immune system. Today 1 in 5 Americans are allergic to food or pollen.

The allergist noted that the Amish come from German-speaking people of Switzerland where the percentage of people with allergies is the same as that of the USA, so he doubted that the difference could be due to genetics. To test his hypothesis he looked to the Amish Community in northern Indiana, where 92% of the children either work on a farm or are exposed regularly to farms.

In his sample of 138 children, only 10 had allergies.

- a. Test the hypothesis that the proportion of Amish people from Northern Indiana is less than that for all Americans.

- b. Interpret the P-Value with a sentence in context.
- c. If we made an error, what type did we make? What are the chances we did? Explain.

Reducing the Probability of Making Type I and Type II Errors

- In Math 127, we will only give three nuggets of information pertaining to managing the probabilities of making errors. The Stat II course covers these ideas in much more detail.
- **First**, the only way to simultaneously reduce the probability of making a Type I or Type II error is to _____. If you have more data, then you will be able to put more faith in your decisions and you will be less likely to make an error.
- **Second**, sometimes the consequences of making a Type I error are more severe than the consequences of making a Type II error (or vice versa). The probabilities move in opposite directions. As $P(\text{Type I Error})$ increases, the $P(\text{Type II Error})$ _____.
- **Third**, if you decide to reject the null, the _____ is the probability you are making a Type I Error. Therefore, the _____ the P-Value, the less likely you are making a mistake.

Blackboard Video 16-G

Two-Tailed Hypothesis Tests

- The hypothesis tests we have run so far have been one-tailed tests.
 - The large vehicles and college cheating examples were right-tailed tests.
 - The high school attendance and allergies examples were left-tailed tests.

Occasionally, our hypotheses are set up so that the alternative is “not equal to”. This happens when we are testing for a change, not an increase or a decrease. When you read a hypothesis test problem, look for context clues to determine if the alternative will be “greater than”, “less than”, or “not equal to”.

Tip: When running a two-tailed test, the P-Value is the area in both tails combined. This doubling of the P-Value is the penalty for running a two-tailed test. If the P-Value is doubled, it will be harder to reject the null – i.e. in general, two-tailed tests need more extreme evidence to reject the null than their respective one-tailed tests.

Example: Fifteen years ago, 30% of Cecil students were on some kind of financial aid. We'd like to test if that proportion has changed since then. A random sample of registration records shows that 53 of the 150 students had received aid. Is this evidence of a real change in the trend?

- a. Run the two-tailed hypothesis test, showing all steps.

- b. Support your conclusion with a 95% confidence interval for the true proportion of all Cecil students receiving financial aid. Comment.

Concluding Remarks

- When you run a right-tailed, “greater than” hypothesis test, always shade from your test statistic to the _____ to get the P-Value.
- When you run a left-tailed, “less than” hypothesis test, always shade from your test statistic to the _____ to get the P-Value.
- When you run a two-tailed, “not equals to” hypothesis test, shade _____ ways to get the P-Value. Shade from the positive-valued test statistic and up, and shade from the negative-valued test statistic and down. Or just find one tail area and _____ it.

Blackboard Video 16-H

Fixed Alpha-Level Testing

- Sometimes when hypothesis testing, we need firm answer with exactly “how much evidence” we need to reject the null hypothesis – perhaps the FDA is approving a new drug by running a double-blind completely randomized experiment.
- The main question: **Exactly** how unusual would the results of the experiment need to be for us to conclude that the new drug is better?
- Sometimes it is better to specify this before the test – this is called the _____ of the test or the _____ of the test.
- It is commonly accepted that _____ or _____ is “rare enough” to reject the null – i.e. if our sample data would only have happened 5% or 1% of the time if the null were true.
- When running a test in this manner, you _____ the null when your P-Value is **less than** α .
- It doesn’t matter if you are running a left-tailed test, a right-tailed test, or a two-tailed test. Reject the null hypothesis if your P-value is _____ α .
- If you reject the null using a significance level of 5%, we say the data are _____ at the 5% level.
- Most times, our hypothesis test problems will not specify an alpha-level. This is because in the real world, most hypothesis tests are performed without a stated alpha-level.
- When running tests without an alpha-level, simply look to the P-Value to make your decision. We know that the smaller the P-Value, the _____ evidence there is in favor of the alternative hypothesis.
- Use common sense and always refer back to the **P-Value Diagram** in Lesson 16-B.
- One disadvantage of not using an alpha-level is that two people can look at the same P-Value and come to different conclusions about a hypothesis test.
- One disadvantage of using a fixed alpha-level test is that a very tiny change in your sample data can lead to a very opposite conclusion.

Example: A statistics intern has tested a new ointment for rashes against the existing ointment. The existing ointment has worked in clinical trials for 69% of users. The P-Value of the test was 0.041.

- a. Write the appropriate hypotheses.
 - b. Your old boss was pretty easy-going and directed you to run your hypothesis tests with a 5% significance level. What conclusion will you make about the ointment. What are the consequences?
 - c. Well, he got fired (we don't know why). We had already collected our data from our volunteers, but hadn't published the results of our hypothesis test. The new boss, she's pretty tough. She requires a 1% level of significance on all tests. What would we now conclude? What are the consequences.

Note: Remember, when asked to run a fixed alpha-level test, don't fret. Just check: If the P-value is _____ alpha, reject the null hypothesis. It's that easy.

Blackboard Video 16-I

Using Confidence Intervals to Run Hypothesis Tests

- Confidence intervals and hypothesis tests are built from the same ideas⁵.
- You can view the values inside a confidence interval as a set of _____ for the true population parameter.

1. Running a **two-tailed** test using a confidence interval.

- You can test, at the significance level α :

$$H_0 : p = p_0$$

$$H_A : p \neq p_0$$

By computing the $(1 - \alpha) \times 100\%$ confidence interval.

- If p_0 falls outside the interval, reject the null hypothesis.
- If you compute a _____ confidence interval, it is effectively like running an _____ hypothesis test.
- If you compute a _____ confidence interval, it is effectively like running an _____ hypothesis test.

Example: In November of 2013, Obama's approval rating stood at 41%, based on a sample of 1000 randomly selected adults.

- a. Make a 95% confidence interval for his approval rating by all U.S. adults.

⁵ Almost correct. Confidence intervals use the standard error while our tests use the standard deviation of the assumed model. For our purposes, the two will usually be close enough that we won't know the difference.

- b. Based on the interval, test the null hypothesis that half of Americans approved of the way he was handling the country at the time. We are looking for evidence that the true proportion differs from 50%. What is the significance level of the test.

2. Running a one-tailed test using a confidence interval.

- You can test, at the significance level $\alpha / 2$:

$$\begin{array}{ll} H_0 : p = p_0 & H_0 : p = p_0 \\ H_A : p < p_0 & \text{or} \\ & H_A : p > p_0 \end{array}$$

by computing the $(1-\alpha) \times 100\%$ confidence interval.

- If p_0 falls outside the interval, reject the null hypothesis.
- If you compute a _____ confidence interval, it is effectively like running an _____ hypothesis test, because we are only concerned with half of the alpha in one tail.
- If you compute a _____ confidence interval, it is effectively like running an _____ hypothesis test, because we are only concerned with half of the alpha in one tail.

Example: Rasmussen Reports in July 2012 published that 26% of Americans felt that their finances were getting better. This report was based on responses from 3,127 adults.

- a. A statistician created a 95% confidence interval for the true proportion of Americans who felt their finances were improving: (24.46% to 27.54%). Can we conclude that more than one-quarter of all Americans feel they are improving? Why?
 - b. What the hypotheses for the one-tailed test?
 - c. What is the significance level of the test if we use our 95% confidence interval to make our conclusion? What is the conclusion?

Blackboard Video 16-J

Lesson 16 Terminology Review and Two Final Quick Examples

- The smaller the P-Value, the more evidence there is in favor of the _____ hypothesis.
- A P-Value is a _____. It is the probability that we drew our sample **if** the null hypothesis is true.
- Data are called _____ if we reject the null hypothesis. It means the data we collected were very unlikely under the assumption that the null hypothesis is true.
- Data are called _____ if we actually take action based on our hypothesis test. You do not need to reject the null to take action, and just because you rejected the null, you do not need to take action. Listen to what the data are telling you!
- **Don't** rely too heavily on fixed-alpha level testing. If your P-Value was 0.049 versus 0.051, at the 5% level of significance, you would make vastly different conclusions. The difference between a P-Value of 0.049 and 0.051 results from a tiny, tiny change in your sample data. Should we be making vastly different conclusions?
- A _____ is rejecting the null when the null is true.
- If we run a fixed-alpha level test, when we specify alpha, we are actually pre-determining the probability of making a Type I Error **that we can live with**.
- If we run a test without a fixed-alpha level, **and** we reject the null hypothesis, then the probability of making a Type I Error is actually the _____.

- Therefore, the smaller the P-Value, the less likely we are making an error when we reject the null hypothesis.
- If the consequences of making a Type I Error are severe, pick a _____ alpha or require a very small _____.
- A _____ is failing to reject the null when the null is false. The probability of this happening is labeled with the symbol _____. This is harder to calculate because the alternative hypothesis represents a multitude of values.
- The only way to reduce the chance of making both errors is to ______. In theory, this is nice, but collecting more data costs money and takes time.

Example: Testing for Lupus is difficult because the signs and symptoms vary considerably from person to person. Recently, a few medical interns created a quick screening test. A patient who tested positive would then be referred to a rheumatologist.

The interns believe there is a false positive rate of 3%. Presume there is a false negative rate of 8%.

a. What will we assume about every patient before going through the initial screening?

b. In words, what are the two hypotheses if we look at this as a hypothesis test?

c. In words, explain what happens if we make a Type I Error.

d. In words, explain what happens if we make a Type II Error.

Example: Gallup reported in November 2013 that 16.6% of all adult Americans are under-employed. This was based on a sample of 2500 people. In March of 2011, the figure stood at a flat 20%. A test was run to determine if this was a statistically significant decline, and the P-Value was 0.00001. Clearly, the researchers concluded there was a decline. Interpret the P-Value in the context of the problem.

Blackboard Video 17-A

Running a Hypothesis Test and Computing a Confidence Interval For the Difference In Two Proportions: Part 1 of 3

Example: The author has a hunch that at Cecil College, a greater percentage of females are smokers when compared to males. This is primarily based on visual evidence from working here the past number of years.

1. We could test that hypothesis by collecting independent samples at Cecil College – one sample from each gender – and determine the sample proportion of each group that were smokers.

We'd be testing the following hypotheses:

Equivalently, we could write it this way:

We don't have to write it both ways. In Lesson 17, we are testing for a _____ between two proportions. "Difference" means _____ . You can see it in the second presentation.

2. We know that unless we get really lucky, the sample proportions will be different. Here's the data. In 120 randomly selected females, we had 30 smokers. In 90 randomly selected males, we had 18 smokers. Summarize it.

3. **The main question:** “How much greater would the sample proportion of female smokers need to be, compared to the sample proportion of male smokers, for us to conclude that indeed, more females smoke than males?”

- Certainly, with our sample sizes, if 50% of the females and 10% of the males smoke, we’d have a no-brainer. We’d conclude, that at Cecil College:

- Certainly, with our sample sizes, if 25% of females smoke and 24.4% of males smoke, we’d also have a no-brainer. We’d conclude, that at Cecil College:

The key idea for comparing two sample proportions is to measure the difference using a standard deviation or a standard error.

4. Things get a little technical in Lesson 17, so we will take it slow. The standard deviation for the difference between two sample proportions is given by the following formula:

This formula assumes we know the values of p_1 and p_2 , which we won’t. It is a theoretical formula without much practicality. For the Cecil smoking example, p_1 would be the proportion of all females that smoke. How would we know that? Same for p_2 with the males.

Blackboard Video 17-B

Running a Hypothesis Test and Computing a Confidence Interval For the Difference In Two Proportions: Part 2 of 3

5. With sample proportions, we will be able to compute the **standard error** for the **difference between two sample proportions** once we collect our data. We substitute the values of \hat{p}_1 and \hat{p}_2 for p_1 and p_2 in the standard deviation formula from 4. Here is the formula:

Using the Cecil College sample data, compute the value of the standard error for the difference in sample proportions:

6. With the standard error value calculated, we can now create a confidence interval for the true difference in the proportion of smokers, females and males, at Cecil College. Here is the confidence interval formula:

Now calculate the 95% confidence interval for the difference Cecil College smoking proportions, females versus males.

7. Interpret the interval with a sentence in context.

Blackboard Video 17-C

Running a Hypothesis Test and Computing a Confidence Interval For the Difference In Two Proportions: Part 3 of 3

8. Because 0% is inside the 95% confidence interval, we can make a particular conclusion. 0% means “0% difference” or “no difference” between the percentage of smokers among the genders.

Because 0% is inside the interval, then we can conclude, statistically-speaking, that at Cecil College:

9. Now let's run the hypothesis test that we started with in Video 17-A.

Recall the hypotheses:

We will first need to convert our data to the test statistic:

- **Here's the tricky part:** Since we are hypothesizing that our proportions are equal (in the null hypothesis), we are going to pool our standard deviations together rather than treat them separately in the denominator of our test statistic formula for z .
-
- Now compute the value of the test statistic for the Cecil College smoking example:
-
9. The tricky part is over. The test statistic follows a Standard Normal model, just like when we ran hypothesis tests for one-proportion. **To get the P-Value**, just shade under the Standard Normal model like we've been doing for our one-proportion tests.

- 10.** Make a decision and write a concluding remark in the context of the problem.

Two-Proportion Problems: Tests and Intervals: Quick Reference

Check Conditions:

1. The variable is categorical and we are working with proportions.
2. Both samples are random or at least unbiased.
3. Both sample sizes are less than 10% of their respective population sizes.
4. We have or expect at least 10 successes and 10 failures in both samples.

Confidence Interval Formula:

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

Hypothesis Test:

$$\begin{aligned} H_0 : p_1 &= p_2 & \text{versus} & \quad H_A : p_1 &> p_2 \\ &&&&\neq \end{aligned}$$

Test Statistic:
$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_1} + \frac{\hat{p}_{\text{pooled}}(1-\hat{p}_{\text{pooled}})}{n_2}}} \quad \text{with}$$

$$\hat{p}_{\text{pooled}} = \frac{x_1 + x_2}{n_1 + n_2}$$

Blackboard Video 17-D

Example: In a sample of 400 Cecil students, 49% had GPA of 3.00+, while in a sample of 700 Harford students, only 46% had GPA of 3.00+. Is this statistical evidence that Cecil students have a higher proportion of 3.00+ students (as opposed to just normal sampling variation)?

- a. Run the appropriate test, performing all steps. Use a 1% level of significance.

- b. Support your test results with a 98% confidence interval for the true difference in the proportion of students with “high” GPAs. Interpret.

Blackboard Video 17-E

Example: An Internet study claims that while 27% of males expect to be financially prepared for retirement, only 18% of the sampled females express that opinion. If the survey was administered to 200 people of each gender, could you conclude that more men expect to be financially ready for retirement?

- a. We need to run a hypothesis test for the difference in attitudes – is this 9% discrepancy a true pattern prevalent in both populations or is it just natural sample-to-sample variation? Run the test at a 0.05 significance level.

- b.** Support your conclusion with a 95% confidence interval for the true difference in the proportions of men and women who are confident to retire comfortably. Interpret.

Blackboard Video 18-A

Modeling the Sample Mean: Part 1 of 2

- Lesson 14 – 17 dealt with _____ variables and sample _____ . For this, there was one set of rules to calculate confidence intervals and run hypothesis tests.
- To finish the course, in Lessons 18 – 21 we now look at _____ variables and sample _____. Of course, there will be a new set of rules⁶, but the ideas are the same. We are still working towards calculating confidence intervals and running hypothesis tests. We need the theoretical model for the sample mean, *if we took repeated samples from our population*, which we won't. But this theoretical model provides the framework for our intervals and tests.

Example: We will investigate IQ scores of Cecil College students, because the idea is familiar and IQ scores follow a Normal model with an established mean of 100 and a known standard deviation of 15 points.

Here is a made-up, but probably realistic, snapshot of some of the IQ scores for all 2500 Cecil College students:

97	103	111	100	129	85	107	110	97	92	113	112
75	94	93	104	94	96	117	105	104	118	119	98
98	58	92	82	72	110	104	128	97	117	92	88
95	105	109	101	83	70	122	86	127	96	119	91
111	84	126	87	89	97	118	100	120	88	99	109
108	112	118	85	94	94	93	111	130	116	112	99
105	108	58	79	103	110	111	112	63	99	79	84
97	75	95	94	110	92	109	107	108	102	92	85
120	108	98	98	94	103	101	91	95	109	75	89
89	104	91	85	109	115	125	99	96	99	112	109
110	103	85	115	87	93	137	120	91	89	95	101
89	99	106	133	105	106	89	103	108	98	110	81
113	93	85	111	85	103	107	95	96	102	118	126
96	95	117	103	87	110	92	107	92	88	127	108
75	88	101	95	96	121	113	108	106	99	114	97
104	124	119	78	110	109	84	84	83	94	99	93
106	93	125	105	119	81	99	94	106	119	98	121
106	86	92	118	103	90	106	105	104	124	109	105
125	95	98	118	103	92	107	108	100	115	121	107
85	111	95	96	114	113	91	88	118	120	108	110
110	96	103	100	84	68	108	123	98	105	100	110
92	101	94	94	112	105	91	99	82	117	96	107
71	113	74	101	78	114	83	104	95	99	105	101
92	110	107	130	65	83	72	89	65	106	97	118
113	81	110	95	83	112	74	111	94	129	102	95
101	107	95	93	83	99	97	81	74	114	119	100
110	109	98	97	124	106	100	110	91	82	93	121
117	88	85	125	112	117	111	107	101	86	100	104
78	80	64	82	134	74	119	89	91	91	114	115
122	127	84	102	103	115	99	118	92	106	72	129
100	115	85	134	85	94	105	108	113	112	101	96
95	105	112	96	105	95	73	94	105	105	92	99

⁶ Past experience indicates that students have difficulty classifying problems as proportion problems or mean problems. Always look at the variable. Are you working with a categorical variable? Proportion problem. Are you measuring or counting up a quantitative variable? Mean problem. Nothing to it.

- We believe that Cecil College students are no different than the general population when it comes to IQ scores, no better, no worse, but who knows until we actually collect our data. We will take a sample of size $n = 100$ randomly selected students this Saturday and administer IQ tests.
- Here are the results from our Saturday IQ test:

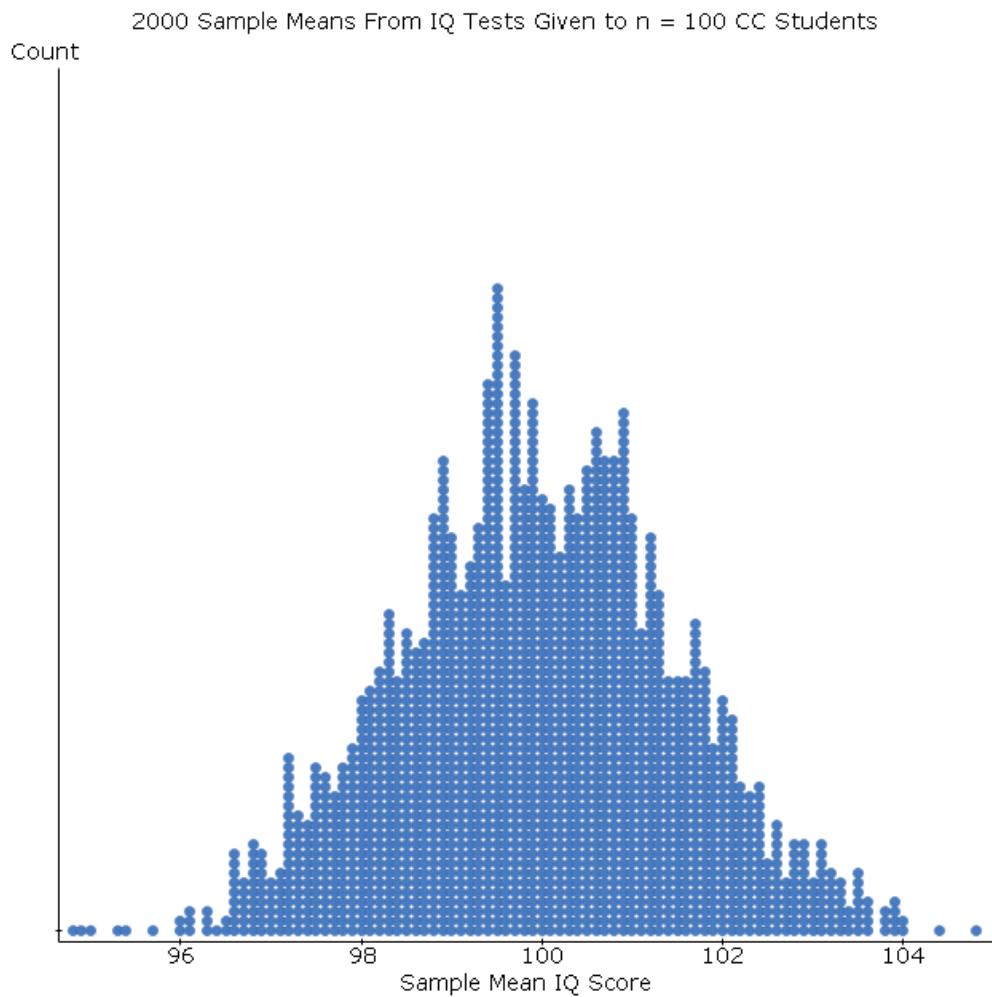
81	69	85	136	130	98	103	95	122	87
95	95	77	110	92	101	94	126	105	105
83	101	83	104	103	94	85	103	98	86
101	130	105	99	104	105	107	90	115	120
97	93	107	94	92	111	98	110	80	98
119	104	74	119	112	105	119	98	104	85
109	92	100	105	95	113	110	108	111	111
80	99	89	93	94	106	118	92	103	79
86	100	99	89	83	104	109	124	75	100
90	82	112	129	108	116	76	112	123	117

- Based on this sample, we could run a hypothesis test to determine if the mean IQ score at Cecil College is any different from that of the general population, $\mu = 100$.
- To run that test, or to compute a confidence interval, we will need the theoretical model for the sample mean, _____. We already determined the theoretical model for the sample proportion, _____, back in Lesson 14. The theoretical model is based on taking repeated samples. Clearly, the IQ scores would change if we randomly selected $n = 100$ new students:

85	87	98	121	120	100	121	97	88	108
81	103	91	99	98	116	78	125	112	102
92	90	96	92	116	98	91	98	84	64
104	120	102	118	88	102	111	108	96	91
113	102	88	106	101	104	86	85	90	99
92	107	123	95	105	101	106	98	126	101
107	122	98	81	128	82	71	105	92	109
107	121	100	105	70	103	98	97	95	92
92	93	89	111	93	96	83	131	102	122
91	109	84	66	68	102	99	71	110	67

- So, we must imagine taking repeated samples of size $n = 100$, each time calculating the sample mean, and then determining the correct theoretical model for the sample mean.
- Once we know the correct model for the sample mean, then in Lessons 19 – 21, we can develop methods to calculate confidence intervals and run hypothesis tests.

- A computer simulation was performed to generate many possible sample means. Samples of size $n = 100$ were generated. IQ scores were simulated. And each time, the sample mean was calculated. Here is a dotplot of the results:



- What theoretical model should be used for the sample mean distribution? _____
- For the theoretical model, what value would you guess for the mean? _____
- The standard deviation of individual IQ scores is _____. Clearly the standard deviation for the theoretical model for the sample mean is much _____ than 15.
- Take our word. The formula for the standard deviation of the sample mean is:

Blackboard Video 18-B

Modeling the Sample Mean: Part 2 of 2

- The theoretical model for the sample mean will be approximately Normal if one of two (or both) requirements is met:

Requirement 1: The population you are sampling from is _____.

Example: IQ scores follow a Normal model. Therefore, no matter what sample size we choose (we did $n = 100$ at Cecil College), the sample mean will also follow a Normal model. We could sample 4 students, 40 students, or 400 students. It doesn't matter, because IQ scores are Normal. Therefore, the sample mean will also be Normal.

Requirement 2: You take a sample size of at least _____.

Note: Do not confuse this with 10 successes and 10 failures. We are now working with quantitative variables, so the requirement is different (even though 10 / 10 or sample size at least 30 do roughly the same thing).

Example: At Cecil College, we'd like to run a hypothesis test to see if the mean age of students has increased since it was last officially announced as 27.5 years old.

- What's the variable and what kind of variable is it?
- What is the likely shape of student ages at Cecil College?
- If we'd like to perform our hypothesis test, how many students will we have to sample?

Official Rules for Modeling the Sample Mean with a Normal Model

1. Check conditions.
 - A. The variable is _____ and we're investigating the _____.
 - B. The sample is _____ or at least _____.
 - C. The sample size should be less than _____ of the population size.
 - D. We are confident the population is _____ or we have taken a sample size of at least _____.
2. If the conditions are met, then a Normal Model is appropriate for the sample mean. We need the mean and standard deviation for this theoretical model:

Central Limit Theorem

- For our purposes, this is more or less a trivia fact (**author's opinion**), but we would be remiss if we didn't mention it officially.
- For most any population (skewed, uniform, bimodal, tri-modal, Normal-ish, etc...), the sample mean will follow a Normal model, as long as you take a random sample (unbiased) of at least size _____. The reasons are a bit beyond the scope of Math 127, but we will show you simulations in class to "prove" it.

Blackboard Video 18-C

Example: The length of a human pregnancy, measured in days, follows a Normal model with a mean of 266 days and a standard deviation of 16 days.

a. Draw the Normal model for individual pregnancies. Label.

b. What proportion of pregnancies last 285 days or longer?

c. What's the probability that any pregnancy lasts between 256 and 276 days?

- d. We take a random sample of 16 pregnancies from the local hospital and would like to answer a few questions about the mean length for the 16 pregnancies. What model will we use? Why can we use that model?
- e. Draw the model for the mean length of 16 randomly selected pregnancies. Label. Comment.

- f. What percentage of the time will 16 randomly selected women have a mean pregnancy length of at least 285 days? Draw. Comment.
- g. What is the probability that a group of 16 randomly selected women have a mean pregnancy length between 256 and 276 days. Draw. Comment.

Blackboard Video 18-D

Example: Assessment records in Cecil County show the mean home value on the books at \$224,000 with a standard deviation of \$65,000 and a very skewed-right shape. Local officials decide to take a sample of 100 homes.

- a. Fully describe the sampling distribution model for the sample mean. Draw. Label. Why can we use this particular model?

- b.** Comment on the population of homes and their assessed values if local officials find the mean assessed value in the sample to be only \$204,000.

Example: There are 33 students in an intro stat course over at the University. From the professor's experience, exams take an average of 5 minutes to grade, with a standard deviation of 5 minutes. Grading time per exam is not Normal.

- a. The professor starts grading at 7:00 p.m. and wants to be finished by 11:00 p.m. That means he needs to finish grading in _____ hours or _____ minutes.
- b. To finish grading in the allotted time, what will he need his "average" grading time to be, per exam?
- c. This is a problem about the distribution of the "average" grading time. Since the sample size exceeds _____, a Normal model will apply. Determine the mean and standard deviation of this Normal model.
- d. Determine the chance he finishes by 11:00 p.m. by determining the probability that the average grading time is less than 7.27 minutes per exam.

Blackboard Video 19-A

Confidence Intervals for Means

- The sampling distribution for the sample mean is based on knowing a key fact that, in practice, we will rarely ever know.
- As long as the conditions are met, \bar{y} will follow an approximate Normal model with:
- In practice, how would we ever really know the value of _____, the standard deviation of the entire population. We are running tests or computing confidence intervals for the mean, _____, which is a measure of the center. If we aren't really sure about the center, how on earth would we precisely know the value of the spread?

Example: We recently read somewhere that the average credit card debt carried by college students is \$3,173. To see if we differ, we'd like to compute a 95% confidence interval for the true mean credit card debt for all Cecil College students. Diagram the population and the sample.

The Big Problem

- We will perform inference on our population, based on our data, based on our collected sample, and based on our sample statistics.
- We only have a “best guess” for the population standard deviation. Our best guess for the population standard deviation _____, is the sample standard deviation, _____.
- If we replace the value of σ with s in our sampling distribution, then \bar{y} no longer has an approximate Normal distribution!
- Thus, everything we learned in Lesson 18 about Normal models will not apply in Lessons 19, 20, and 21.
- In Lesson 18, we were working in a sterilized environment – all of the values of the population parameters were known. We knew the population mean, μ . We knew the population standard deviation, σ . In the real world, when working with sample data, we won’t know these values, so we must approximate them with sample statistics. When we do this, we need a new model to serve as the basis for our confidence intervals and hypothesis tests.

The Big Solution

- When we collect our sample data, we will work with the **standard error** of the sample mean, rather than the standard deviation of the sample mean:

Real World, With Real Data:

Theoretical, Sterilized World:

Gossett

- The statistician credited with discovering the new model we will use for working with means was named William Gossett. He made his important discoveries during the early _____, but not without controversy.
- Gossett worked for _____ brewery. He tested beer using statistical methods, but frequently was making errors because he was using the wrong model.
- Gossett discovered that, especially when working with small sample sizes, using the Normal model was inappropriate after substituting s in for σ .
- Rather, a model that was also _____, but with a bit more area in the tails when compared to the Standard Normal model, was called for.
- When using this new model, confidence intervals tend to be a little bit _____ on purpose.
- When using this new model, hypothesis tests generated P-Values that were a little bit _____ on purpose.
- Guinness felt this statistical discovery was a trade secret, to be kept in-house, but Gossett convinced management that his discovery would be of no practical use to competing breweries.
- Finally, Gossett was allowed to publish his results, but under the pseudonym “Student”.
- Therefore, his noteworthy achievement does not carry his name, and Gossett’s distribution is called the _____.

Important

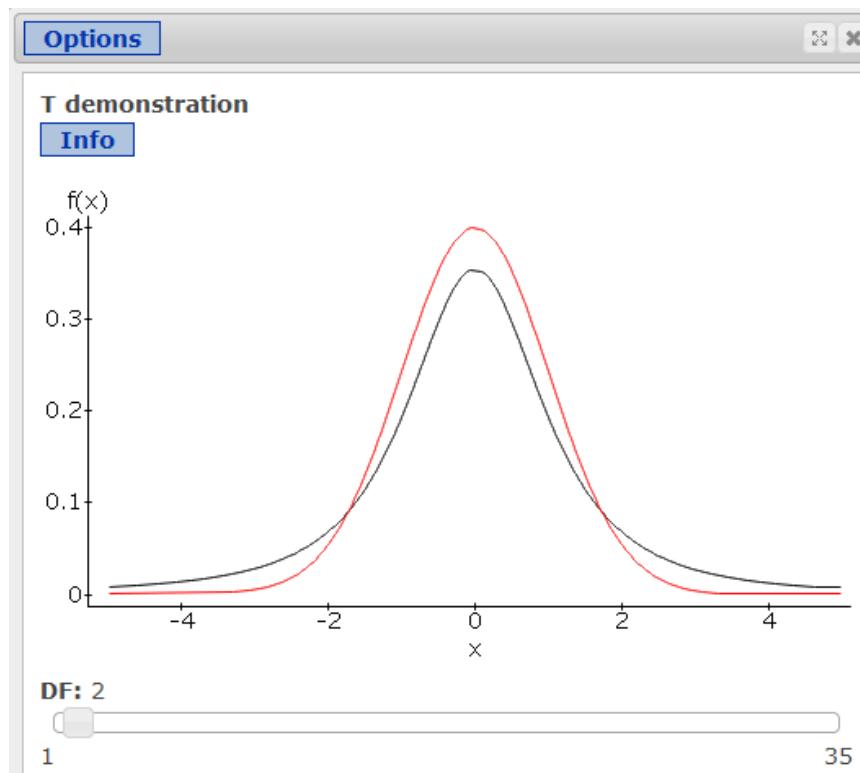
- When you are performing inference on **proportions**, use the _____ model, like we did in Lessons 14 – 17.
- When you are performing inference on **means**, Lessons 19 – 21, use the _____ model.

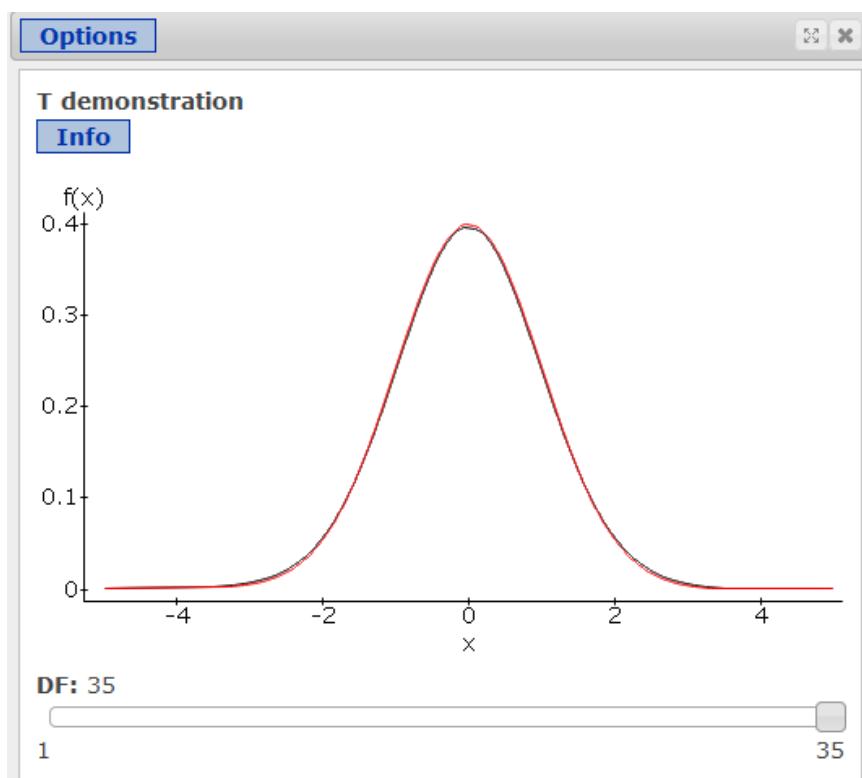
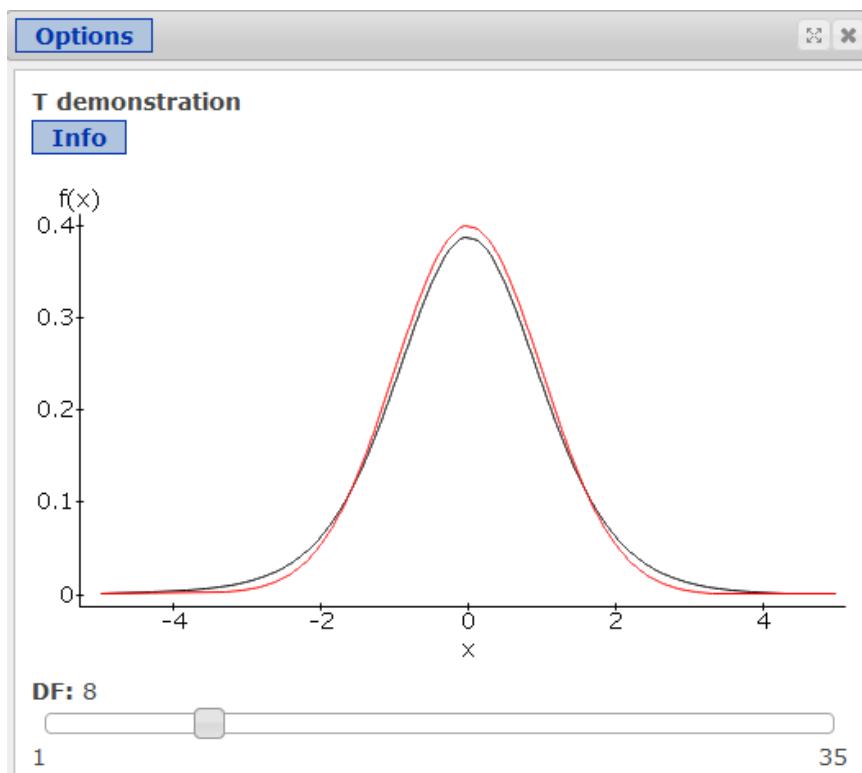
Blackboard Video 19-B

Some Facts about the Student's t Distribution

- The shape of the model is _____, but compared to the Standard Normal model, there is more area in the tails.
- The larger your _____, the closer the Student's t model will be to the Standard Normal model. In fact, when the sample size exceeds about 100, there is very little difference between the Student's t and the Standard Normal model.
- For very small sample sizes, the Student's t model will have **much** more area in its tails.
- The Student's t is a family of models, one for each sample size. Officially, we name each one by its _____, but it's all based on sample size. Typically, degrees of freedom = $n - 1$.

Example: Observe the difference between the Standard Normal model and the Student's t model for varying degrees of freedom.





Developing Confidence Intervals and Hypothesis Tests for Means (One-Sample)

1. Check Conditions:

- A.** The variable is quantitative and we are working with means.
- B.** The sample is random, or at least unbiased.
- C.** The sample size n is less than 10% of the population size N .
- D.** The population we are sampling from is Normal or we have a sample size of at least 30.

2. Gossett Discovered:

3. Confidence Interval Formula for a Population Mean μ (One-Sample):

Blackboard Video 19-C

Finding Critical t Values on StatCrunch

Example: Say we'd like to create a 95% confidence interval for a population mean and we collected a sample of size 15.

- a. The confidence interval formula $\bar{y} \pm t\left(\frac{s}{\sqrt{n}}\right)$ requires the value of t . Find it on StatCrunch.
- b. If we incorrectly used the Standard Normal model to create a 95% confidence interval, give the incorrect critical value z . Notice the t value is correctly bigger.

Example: Say we'd like to create a 99% confidence interval for a population mean and we collected a sample of size 7.

- a. The confidence interval formula $\bar{y} \pm t\left(\frac{s}{\sqrt{n}}\right)$ requires the value of t . Find it on StatCrunch.

- b. If we incorrectly used the Standard Normal model to create a 99% confidence interval, give the incorrect critical value z . Notice the t value is correctly bigger.

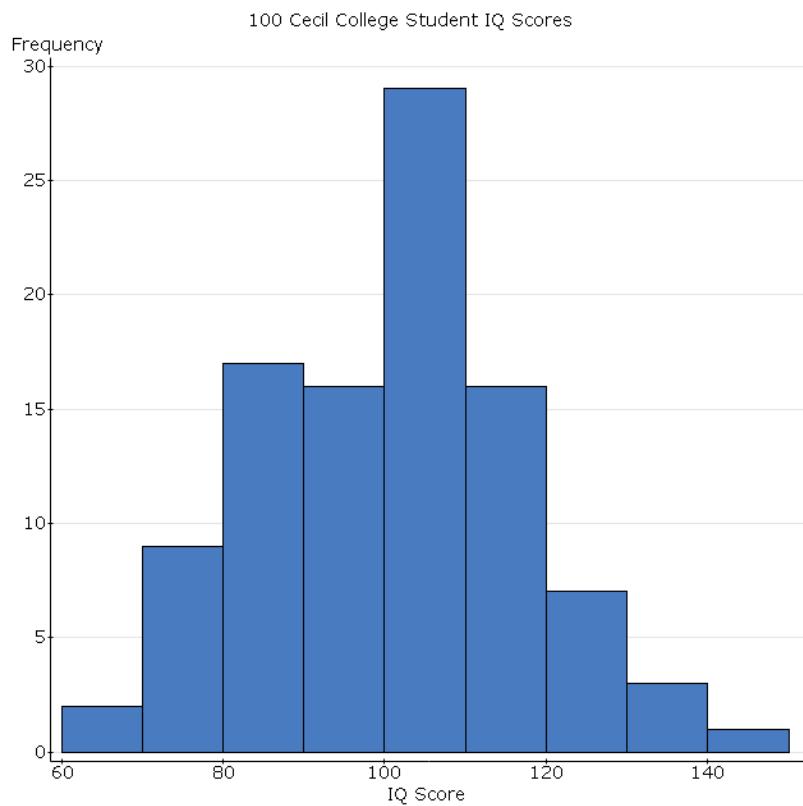
Blackboard Video 19-D

Computing Confidence Intervals for a Population Mean

Example: 100 Cecil College students were recruited to take an IQ test to determine if our students differ from the general population. The data is printed below:

107	101	113	125	91	89	103	140	96	86
76	102	71	104	89	77	105	102	107	115
102	110	117	113	92	87	101	76	124	109
88	77	111	83	111	135	136	88	114	96
108	104	98	108	94	122	84	134	94	100
89	122	108	79	65	114	86	117	126	128
127	95	105	93	116	93	89	105	75	86
65	103	105	112	112	95	100	99	108	85
113	86	100	86	116	107	91	100	92	85
91	104	78	109	107	81	93	73	115	102

- The statistician ran the appropriate summary statistics and created a histogram:



Summary statistics:

Column	n	Mean	Std. dev.	Median	Min	Max	Q1	Q3
IQ Score	100	100.46	15.959178	101.5	65	140	89	111

- a. To proceed with our confidence interval for the true average IQ score for all Cecil College students, certain requirements must be met. Are they? Explain.
 - b. How many degrees of freedom? Determine the critical t value for a 95% confidence interval for the true mean.
 - c. Using the formula, calculate the 95% confidence interval.

- d. Interpret the interval with a sentence in context.
- e. Do we have any statistical evidence that Cecil College student IQs differ from the generally agreed upon average of 100 points? Why or why not?
- f. What is the margin of error? What factors are influencing its value?

Blackboard Video 19-E

Example: Professor Kupe bowled on a men's league team last year and told his team captain that he was a 200 average bowler. After the 96 game season, do we have statistical evidence that he might have been inflating his abilities to "sound good" to his captain? Here is the summarized data:

Summary statistics:

Column	n	Mean	Std. dev.	Median	Min	Max	Q1	Q3
Scores	96	188.05208	30.389877	182	129	257	165	214

- a. Calculate a 99% confidence interval for Professor Kupe's true mean bowling score, assuming last season was an unbiased representation of his abilities. Assume all conditions are met.

b. Comment on Professor Kupe's claim. What **do** we have evidence of?

Blackboard Video 19-F

Determining Sample Size When Working With Means

- Before engaging in a statistical study to investigate a population, care should be taken to collect a sample size that will meet your needs. A few things to consider:
 1. What _____ can you live with? If your interval is too wide, it will be difficult to make useful decisions with it.
 2. What _____ do you require? The main two choices are 95% and 99%. If there are severe consequences when creating intervals that do not capture the parameter we are estimating, then be more confident.
- The sample size formula comes from algebraically manipulating the margin of error formula:
- **Problems!** For one, we won't know the value of s until we collect our data, which we haven't done yet! Also, the value of t depends on the degrees of freedom, which depends on the sample size, which we are in the process of determining!
- **Solution, Part 1!** Use the equivalent critical z value in place of t . Reminder:

For 90% confidence, use $z = \underline{\hspace{2cm}}$ For 95%, $z = \underline{\hspace{2cm}}$

For 98%, $z = \underline{\hspace{2cm}}$ For 99%, $z = \underline{\hspace{2cm}}$

- **Solution, Part 2!** Come up with an educated guess for the value of the standard deviation. Don't make a wild guess. Usually, there will be a previous study or common-sense value we can use.

Formula to Determine Sample Size When Working With Means:

Example: Say we'd like to estimate the average IQ for Cecil College students. We know that in general, IQs follow a normal curve with a mean of 100 and a standard deviation of 15. We'd like to be 99% confident and we'd like to be within 2 points of the true mean. How many students should be sampled?

Example: Walmart managers want to estimate the average receipt amount for their competitor Aldi, so they will attempt to collect data from customers in the parking lot. They want to get within \$3 of the true mean amount spent and will go with 95% confidence. They know that their own standard deviation for groceries is \$32, so they will use that value in their calculation. How many customers should they attempt to survey?

Blackboard Video 20-A

Hypothesis Tests For Means

Official Steps to Run a One-Sample t -Test for a Mean

1. Write the hypotheses.

Null Hypothesis:

Alternative Hypothesis:

2. Check conditions.

The variable is _____, and we are working with means.

The sample is _____ or at least _____.

The sample size n is less than _____ of the population size N .

The sample size n is at least _____ or there is strong evidence the population is _____. Check a histogram and a QQ-Plot of the sample data.

3. Run the test. **First**, convert your data to the test statistic:

Second, get the P-Value by shading under the Student's t distribution in the H_A direction:

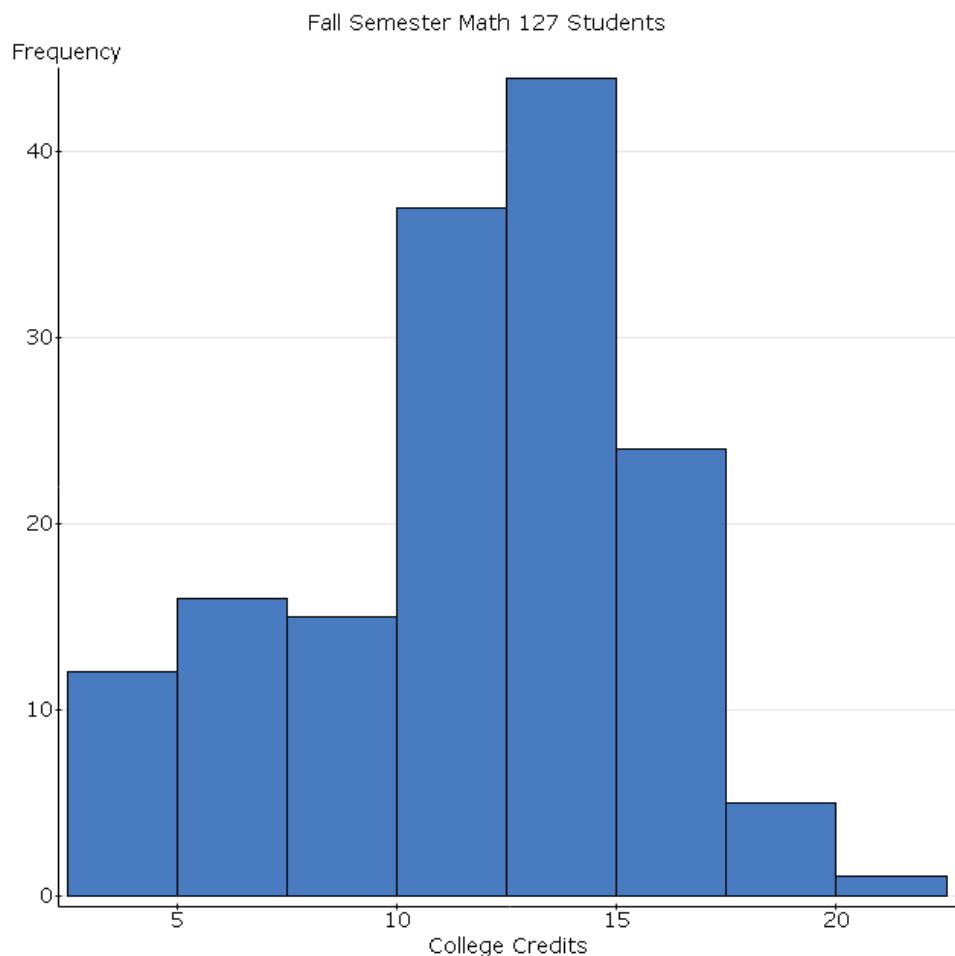
4. Make your decision. If the P-Value is “small”, reject the null in favor of the alternative. If the P-Value is “large”, fail to reject the null hypothesis.
5. Write a concluding remark, based on your decision, in the context of the problem.

Example: One-hundred fifty-four Math 127 students filled out a survey at the beginning of a recent fall semester, and one question asked was “***How many credits are you carrying this semester?***” Presuming Math 127 students are a representative sample of all Cecil College students when it comes to credit-load, do we have **very strong** evidence that Cecil students are, on average, below full-time status?

Full-time status is 12-credit hours. The summarized data is below. Run the *t*-test, showing all steps.

Summary statistics:

Column	n	Mean	Std. dev.	Range	Min	Max	Q1	Q3	IQR	10th Per.	90th Per.
Credits	154	11.565	3.749	18	4	22	8	14	6	7	16



1. Read and understand the context of the problem. Determine if the variable is categorical or quantitative and if you're working with proportions or means. Determine if you are dealing with one or two samples.
2. Write down the hypotheses.
3. Check the appropriate conditions before proceeding.
4. Based on the collected sample data, convert your sample mean to a test statistic.

5. Determine the P-Value.

6. If the P-Value is small, then you'll reject the null hypothesis in favor of the alternative.

7. Write a concluding remark in the context of the problem.

Question: If we made an error, what type did we make and what would that mean in the context of the problem?

Blackboard Video 20-B

Example: Pretend we work for Consumer Reports magazine, and we plan to recommend those Continental DWS tires with the 50,000 mile warranty next month, unless our sample of size $n = 20$ tires convinces us that Continental is not truthful in their claim. We will test the population mean tire tread.

- a. What conditions must be met to proceed, given our limited information?
 - b. Write the hypotheses.
 - c. One member of our team claims that we don't have to run the test because the sample mean from our sample of $n = 20$ tires was $\bar{y} = 49,464$. "The mean is under 50,000 miles, case closed," he says. Explain the faulty logic.

d. Also from the sample, we had a standard deviation of 3,322. Finish running the hypothesis test, and make a conclusion in context. Use $\alpha = 0.01$.

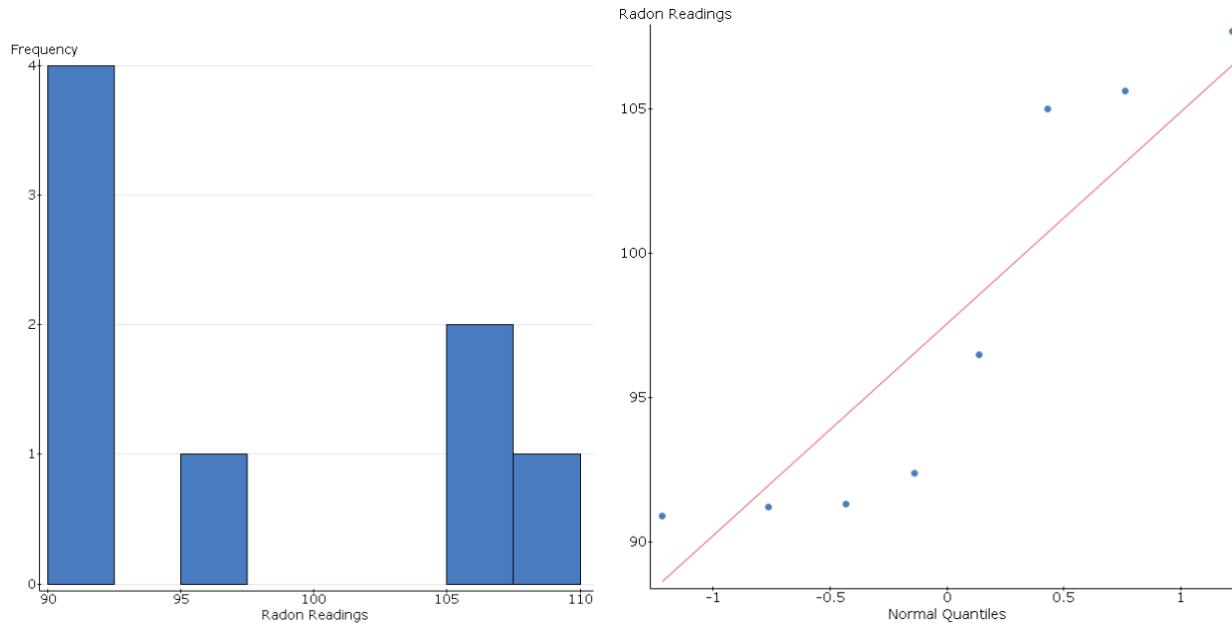
e. Interpret the P-Value with a sentence in context.

- f. Support your conclusion with a 99% confidence interval.
- g. What is the margin of error from the interval?
- h. If Consumer Reports would like to get that margin of error down to 500 miles, to really see more precisely just how long these tires last on average, what sample size would be required?

Blackboard Video 20-C

Example: An engineer tested 8 radon detectors of a certain model for quality control. Each was exposed to 100 pCi/L of radon in a controlled environment. The resulting radon readings were recorded in the table and displayed in a histogram and QQ plot:

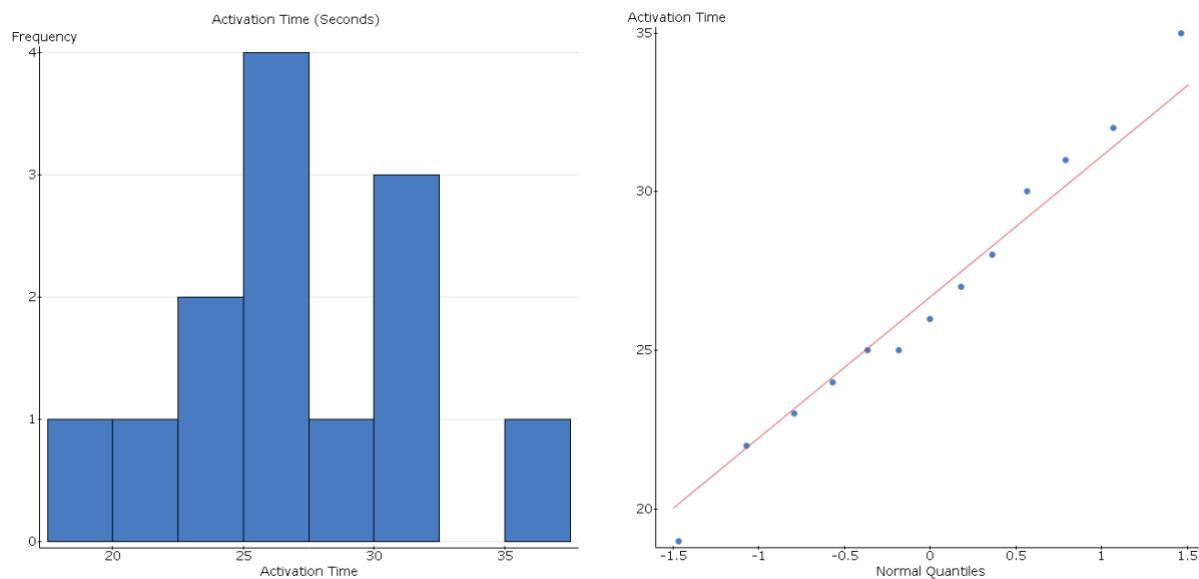
105.6	90.9	91.2	96.5	91.3	105.0	107.7	92.4
-------	------	------	------	------	-------	-------	------



- The engineer would like to run a test to determine if there is compelling evidence that the radon detectors are faulty. Why should you halt the process?
- What suggestion could you give to the engineer?

Example: An office-building-fire-sprinkler supplier runs tests to determine if their sprinklers activate in under 30 seconds, on average. Thirteen sprinklers were tested this month. Can the company still maintain their “under 30 seconds” claim in their advertising? The data and graphs are below:

27	35	22	26	23	32	30	31	25	25	28	19	24
----	----	----	----	----	----	----	----	----	----	----	----	----



Run the appropriate hypothesis test, showing all steps.

Blackboard Video 21-A

Two-Sample Mean Methods

- In Lesson 20, we compared one-sample of data, summarized with a sample mean, to a fixed parameter value.

Example: We tested for evidence that the population mean tire tread was less than 50,000 miles. We tested for evidence that the population mean time for a fire sprinkler to activate was under 30 seconds. Both of these are one-sample mean problems.

- In Lesson 21, we now compare two samples to each other. For example, do Ivy League students have a higher average IQ than Cecil College students? Is there a difference in the mean weights between NCAA football players in Division I and Division III? Is there a statistically significant difference between husbands' and wives' emotional intelligence score? Can Cecil College students do better on the math placement test after attending math boot camp for one week?
- There are two scenarios we must look at in Lesson 21: _____ samples and _____ samples.
- **Independent samples:** The data collected in first sample have no impact on the data collected in the second sample.
- **Examples of independent samples:** Ivy League versus Cecil College IQ scores. NCAA football player weights, Division I versus Division III.
- **Dependent samples:** The data collected in the first sample is paired with the data collected in the second sample.
- **Examples of dependent samples:** Husbands' EQ versus wives' EQ. Before and after math placement test score.
- The methods to test independent samples or dependent samples are quite different. When reading hypothesis test problems, follow this pathway:
 1. Means or proportions?
 2. Then, one or two samples?
 3. Then, if means, independent or dependent samples?

Official Steps: Two-Sample t Test (Independent Samples)

1. Write the hypotheses.

Null Hypothesis:

Alternative Hypothesis:

2. Check conditions.

The variable is _____, and we are comparing means from two independent samples.

Both samples are _____ or at least _____.

Both sample sizes n_1 and n_2 are less than _____ of their respective population sizes.

Both sample sizes n_1 and n_2 are least _____ or there is strong evidence that both populations are _____. Check histograms and QQ-Plots of the sample data.

3. Run the test. **First**, convert your data to the test statistic⁷:

⁷ In this notebook, you will be given the degrees of freedom, because the formula is quite complicated. In class, when we run the two-sample t test, StatCrunch will automatically provide the degrees of freedom.

Second, get the P-Value by shading under the Student's t distribution in the H_A direction:

4. Make your decision. If the P-Value is “small”, reject the null in favor of the alternative. If the P-Value is “large”, fail to reject the null hypothesis.
5. Write a concluding remark, based on your decision, in the context of the problem.

Note: We could write the hypotheses in this equivalent manner:

- The two-sample t test is a test for a **difference** in population means. Difference means **subtract**. The author thinks it's clearer to write the hypotheses as we did on the previous page, but either is correct. You do not have to write them in two ways.

Blackboard Video 21-B

Example: Does college provide students with a beneficial education? Important skills such as critical thinking and reading comprehension are measured by the CCLA or CLA test.

Independent groups of students take the test – a group of incoming first-time full-time students and a group of graduating students. The idea behind the test is simple: if college provided a worthwhile education, then the graduates ought to have a higher mean score on this exam.

Data was found on the internet for Fall 2007 – Spring 2008 for Howard Community College. The results listed below are for the “Analytic Writing Tasks” portion of the test.

	<i>n</i>	Mean	SD
1st Year Students	65	1026	158
Exiting Students	66	1101	133

1. Read and understand the context of the problem. Determine if the variable is categorical or quantitative and if you’re working with proportions or means. Determine if you are dealing with one or two samples.

2. Write down the hypotheses.

3. Check the appropriate conditions before proceeding.
 4. Based on the collected sample data, convert your data to the test statistic. The degrees of freedom are $df = 124.699$.
 5. Determine the P-Value.

6. If the P-Value is small, then you'll reject the null hypothesis in favor of the alternative.

7. Write a concluding remark in the context of the problem.

Question: If we made an error, what type of error did we make? What are the chances we made an error? Explain what is really happening at Howard CC if we did make an error.

Blackboard Video 21-C

Two-Sample Confidence Interval for the Difference in Means

- We can support our hypothesis tests by calculating confidence intervals for a range of plausible values for the true difference in means.

Example: In the CCLA test example, we rejected the null and concluded that grads do better on average than first year students. The next logical question might be, “Well, by how much?” Our best guess would be:

Formula for a Two-Sample t Interval

Note 1: In this notebook, we will give you the degrees of freedom because the formula is complicated. When using StatCrunch directly, df will be calculated for you.

Note 2: When calculating a confidence interval for the difference in two means (or proportions), we look to see if _____ is inside the interval.

If _____ is not included in the interval, then we have statistical evidence that there is a difference in population means (or proportions).

If _____ is inside the interval, then we do not have evidence for a difference in population means (or proportions).

Example: We concluded that graduates do better on the CCLA. By how much. Give a range of plausible values with a 95% confidence interval for the true difference. Degrees of freedom will be 124.699. Recall:

	<i>n</i>	Mean	SD
1st Year Students	65	1026	158
Exiting Students	66	1101	133

Example: Long ago, Math 127 switched textbooks, before moving to this self-authored workbook.

The Old Book was believed to be easier on the homework, when compared to the New Book.

Professor Kupe downloaded all the homework scores from Spring 2011 and compared them to all his homework scores from Spring 2010. We will treat the two semesters as representative samples of all students taking the course from the Old Book versus the New Book.

All the scores of “0” were deleted from both data sets. Here are summary stats:

Variable	N	Mean	Std. Dev	Q1	Median	Q3	Maximum
New Book	1123	87.512	19.503	83.610	95.830	100.000	106.670
Old Book	458	90.820	17.868	91.693	98.150	100.000	100.000

- a. Run the appropriate hypothesis test to determine if we have evidence that the New Book has a lower online homework score (indicating that it is indeed harder than the Old Book). Show all steps. The degrees of freedom for this problem are 920.48.
- b. Support your hypothesis test with a 99% two-sample t interval for the true difference in means. Interpret.

Blackboard Video 21-D

Testing for a Difference with Dependent Samples

- When working with dependent samples, you will have pairs of data values. Simply take the difference for each pair, and run a one-sample t test on the differences or compute a one-sample t interval for the differences.

Example: Assume for Math 127 that take-home quiz 1 and take-home quiz 2 are of equal difficulty. Your instructor thinks grades for the second quiz ought to be higher since students can adapt to the teaching style and know what to expect.

Below is a random sample of 30 students from a recent Spring semester.

Quiz 2	Quiz 1	Difference
83.5	85	-1.5
64	59	5
96	99	-3
68	51	17
83	87.5	-4.5
100	94	6
99.5	95.5	4
81	77.5	3.5
100	98	2
97	93.5	3.5
80	87.5	-7.5
90.5	86.5	4
95	93	2
93	90	3
94	96	-2
97	97	0
94.5	94	0.5
91.5	97	-5.5
71	63.5	7.5
90.5	81	9.5
72.5	83.5	-11
90	86	4
78.5	87.5	-9
84	89	-5
90.5	93	-2.5
79.5	87	-7.5
81	73.5	7.5
91	91	0
91	89.5	1.5
71	77	-6

- We have two samples. Explain why these are dependent samples.
- We will run one-sample t test on the column of differences. What are the appropriate hypotheses?
- We trust by now you could calculate the appropriate summary statistics using StatCrunch. Give the sample mean and sample standard deviation.

- d. Finish running the one-sample t test on the column of differences. Make a conclusion in context.

Example: Metals in drinking water affect taste, and very high concentrations can be harmful. One river which supplies drinking water to a town had six river locations selected. The zinc concentration (mg/L) was measured at the surface and at the bottom of the river. The six pairs of data are provided in the table. Do we have evidence that the true average concentration in the bottom water exceeds that of the surface water?

Zinc at Bottom	0.440	0.256	0.667	0.541	0.707	0.716
Zinc at Surface	0.415	0.238	0.490	0.420	0.632	0.609
Difference						

a. Run the appropriate hypothesis test, showing all steps.

b. Since the difference is statistically significant, support your test results with a 95% confidence interval for the true mean difference in zinc concentrations.

Math 127 Unit III Checklist

- I understand the sampling distribution for a sample proportion and can use the theory to create the appropriate Normal model.
- I can create a one-proportion or a two-proportion confidence interval.
- I can run a one-proportion or two-proportion hypothesis test, showing all steps.
- I understand the sampling distribution for a sample mean and can use the theory to create the appropriate Normal model.
- I also understand that for inference about means, I must use the Student's t model because the population standard deviation will be unknown.
- I can create a one-sample or a two-sample t interval for a mean.
- I run a one-sample or a two-sample t test for a mean.
- I can determine sample size for an upcoming study when working with proportions or means.
- I can differentiate between independent samples and dependent samples when working with two-sample mean problems.
- I understand the types of errors I can make when running a hypothesis test and understand how to interpret their ramifications.
- I understand how to interpret a P-Value, how to interpret a confidence interval, how to make conclusions for hypothesis tests, all in the context of the problem.

