

Math 243 Statistics I

8-week In-Person

Course Packet

Instructor: Cara Lee

Contents

- Video Lecture Notes, pages 1-72
- Graded Problem Sets, pages 73-84

Blank page

Intro to Statistics

Definitions:

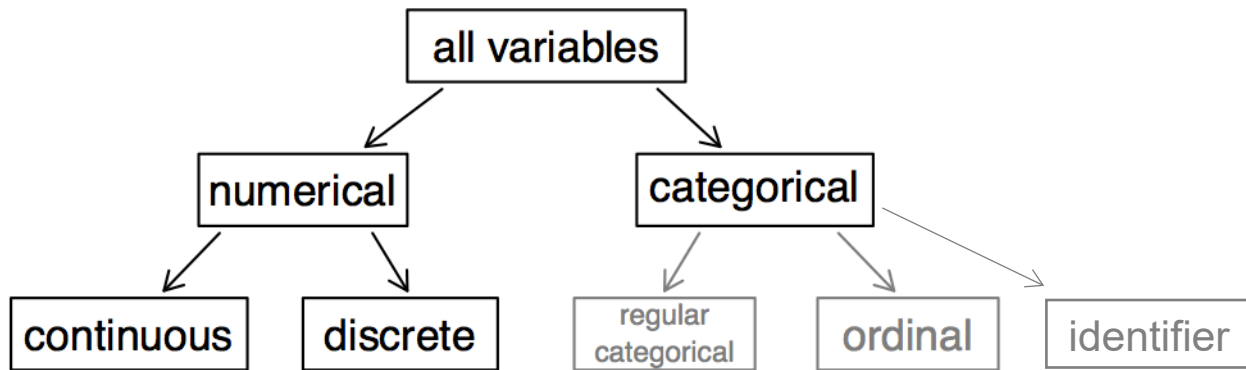
- **Statistics –**
The science of collecting, organizing, summarizing, and drawing conclusions from data.
- **Data –**
Any collection of numbers, characters (words), images, or other items that provide information.
 - **Population Data –** *ALL of the data*
 - **Sample Data –** *A portion of the population data*

Example 1: Draw a picture representing the concepts of **sample** and **population**. Show an example for collecting the average height of a PCC student.

- **Descriptive Statistics –**
Analyzing numerical values that describe and summarize data.
- **Inferential Statistics –**
Generalizing from samples to populations

Reading Data**Data Basics – Types of Variables**

Variables – The characteristics being recorded or measured



Example: Use the data below to explore the ideas of cases and variables.

ID	Gender	Smoke	Award	Exercise	TV	GPA	Pulse	Birth
1	M	NO	OLYMPIC	10	1	3.13	54	4
2	F	YES	ACADEMY	4	7	2.5	66	2
3	M	NO	NOBEL	14	5	2.55	130	1
4	M	NO	NOBEL	3	1	3.1	78	1
5	F	NO	NOBEL	3	3	2.7	40	1
6	F	NO	NOBEL	5	4	3.2	80	2
7	F	NO	OLYMPIC	10	10	2.77	94	1
8	M	NO	OLYMPIC	13	8	3.3	77	1
9	F	NO	NOBEL	3	6	2.8	60	2
10	F	NO	NOBEL	12	1	3.7	94	8

Analyzing Data**Notation**

Type of Data	Quantity of Interest	Population Parameter	Sample Statistic
Categorical (yes/no)	Proportion	p	\hat{p}
Numerical	Mean	μ	\bar{x}

Overview of Data Collection Principles

Example: Researchers are interested in how many people in a city support a new recycling law. Draw and label a picture to represent the population, the sample, the parameter, and the statistic.

Example: Researchers are interested on the average days of shelf life for an apple. Draw and label a picture to represent the population, the sample, the parameter, and the statistic.

Observational Studies & Experiments

Anecdotal Evidence: A single or small number of results, often unusual cases that are not representative of the population.

Observational Studies: In an observational study, researchers gather data without interacting with the subjects.

Retrospective	Prospective

Randomized Experiments: In a controlled, randomized experiment, researchers assign treatments to groups of subjects and measure a response variable.

Practice: Match the description with the proper term.

<p>a. A group of disabled women aged 65 and older were tracked for several years, ending in 2010. Those who had a vitamin B12 deficiency were found to be twice as likely to suffer severe depression as those who did not.</p>	<p>1. Retrospective Observational Study</p>
<p>b. Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred people who identify as men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.</p>	<p>2. Prospective Observational Study</p>
<p>c. Researchers who examined health records of thousands of males found that men who died of myocardial infarction (heart attack) tended to be shorter than men who did not.</p>	<p>3. Anecdotal Evidence</p>
<p>d. A doctor worked with two patients whose depression was cured with vitamin B12 injections.</p>	<p>4. Experiment</p>

e. In part c above, is it correct to conclude that shorter men are at higher risk of dying from a heart attack? Could there be a **lurking or confounding variable**?

f. In which of the above situations can we infer causation? Why?

Sampling Methods

Example: We want to survey PCC students on how much they pay for housing per month. Give an example for each type of sampling.

Method	Description	Example
Census		
Simple Random Sample		
Stratified		
Cluster		
Systematic		
Multistage		

Biased Methods

Bias – Any systematic failure of a sampling method to represent the population. A sample is biased if it does not represent the true population. There is no way to fix biased data so it is better to design a good survey to begin with.

Method	Description	Example
Voluntary or Self-Selected Sampling		
Convenience Sampling		

Experiments

Example: Chia seeds and weight loss. Chia Pets - those terra-cotta figurines that sprout fuzzy green hair - made the chia plant a household name. But chia has gained an entirely new reputation as a diet supplement. In one 2009 study, a team of researchers recruited 38 men and divided them randomly into two groups: treatment or control. They also recruited 38 women, and they randomly placed half of these participants into the treatment group and the other half into the control group. One group was given 25 grams of chia seeds twice a day, and the other was given a placebo. The subjects volunteered to be a part of the study. After 12 weeks, the scientists found no significant difference between the groups in appetite or weight loss.

Vocabulary

a. Describe the **treatment group(s)** (group(s) receiving each treatment/factor)

b. Describe the **control group** (group not receiving a treatment)

Has **blinding** been used? (**single** or **double**) Has a **placebo** been used?

c. Has **random assignment** been used?

d. Has **blocking** been used? If so, what is the blocking variable?

e. What is the **response variable** (what was measured, including units)?

f. Can we generalize the conclusion to the population at large (conclude **causation**)?

Graphs of Numerical Data

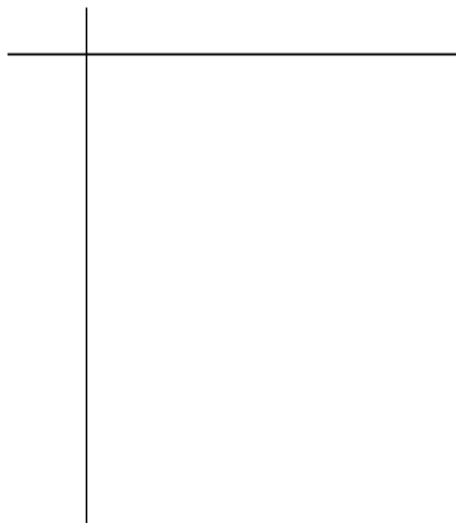
Example: Here is a set of 15 exam scores for a fictional MTH 243 Statistics class at PCC.

31 62 65 70 76 81 82 82 87 88 89 94 95 98 100

Draw a **dot plot** for this data.

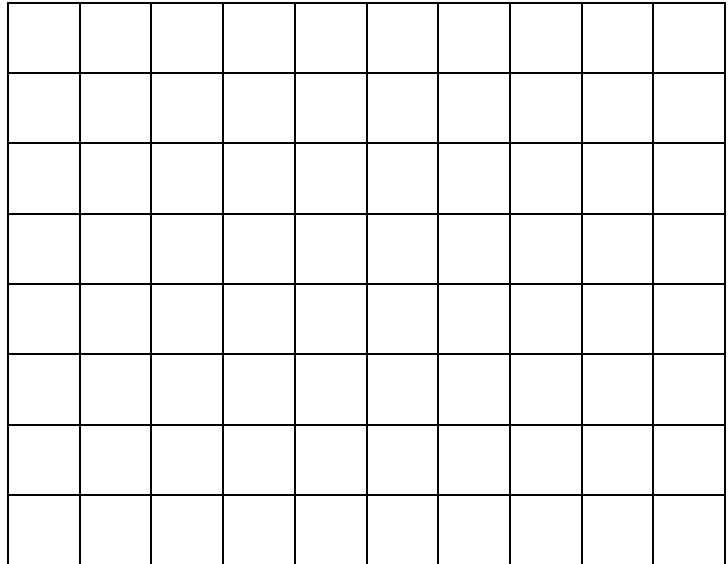


Draw a **stem-and-leaf plot** using the tens digits as the stem and the ones digits as the leaves.

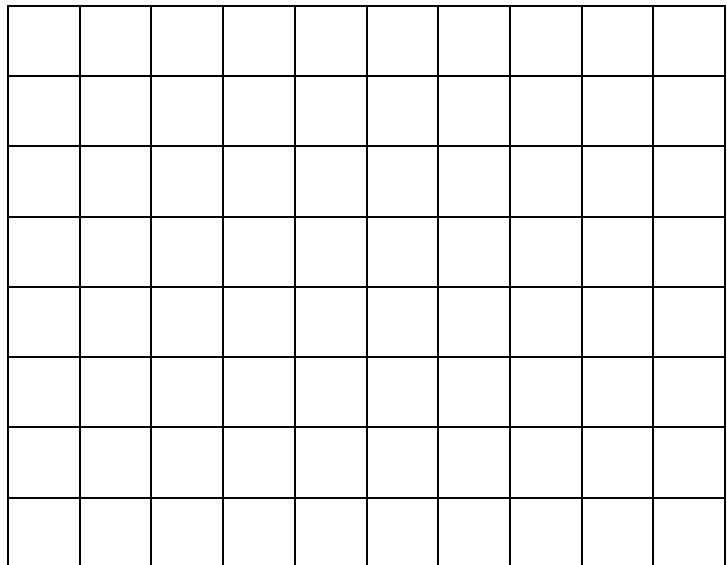


31 62 65 70 76 81 82 82 87 88 89 94 95 98 100

Sketch the corresponding **histogram** for this data using a bin width of 10. Scale and label your graph appropriately.



Sketch the corresponding **relative frequency histogram** for this data.



Describing Graphs of Numerical Data**SHAPE**

Symmetric

skewed left

skewed right

MODE

no mode

unimodal

bimodal

multimodal

OUTLIERS

no outliers

outliers

Mean & Standard Deviation

The *mean* is the average of the data. We calculate this by adding up all values and dividing by the number of values.

Population mean = μ Sample mean = \bar{x}

$$\bar{x} = \frac{\sum x}{n}$$

Example: Calculate the mean of the three data sets:

Data Set A

3 8 12 15 18

Data Set B

3 8 12 15 18 20

Data Set C

3 8 12 15 18 205

For symmetric data, we want to use the mean and standard deviation to describe the center and spread of the data. We would not use these to describe skewed data since the values themselves are skewed.

The difference between one value in a data set and the mean is referred to as a **deviation**. The **standard deviation** is the average distance away from the mean.

- Data set: 1 2 4 5

Because negative distances don't exist, we look at the squares of each of the distances.

To take an average we divide by the sample size minus one, this is called **degrees of freedom**.

The sum of these squared deviations divided by the number of values minus 1 is referred to as **variance**.

The square root of the variance is referred to as the **standard deviation**. It's used to summarize the spread of symmetric distributions and takes into account how far each value is from the mean

- | | | |
|-----------------------|---|---|
| • Deviation: | $x_i - \bar{x}$ | |
| • Variance: | $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ | (population variance σ^2) |
| • Standard Deviation: | $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$ | (population standard deviation σ) |

Example: Calculate the standard deviation for the following data

set: Data Set A

3 8 12 15 18

Example: If Data Set A describes the age of 5 randomly selected people visiting the Oregon Zoo. Describe the center and spread for this data set given that the data is symmetric.

Median & Interquartile Range

The *median* value of a set of data is the “middle value” and divides the data into two equal halves.

Two Cases:

- Odd # of values (Data Set A)

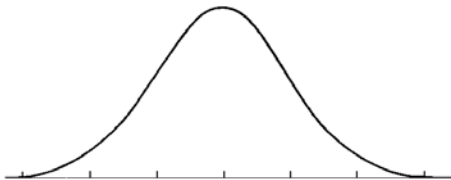
3 8 12 15 18

- Even # of values (Data Set B)

3 8 12 15 18 20

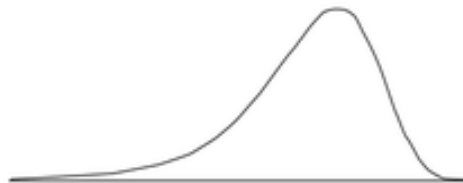
How do the mean, median and mode relate to the shape of the distribution?

Symmetric Distributions

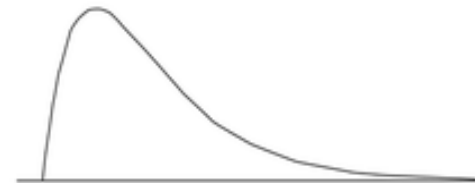


Skewed Distributions

Skewed Left



Skewed Right



For skewed data we want to use the median and interquartile range to describe the center and spread of the data since it essentially ignores the value of any outliers.

The median (Q_2) is the middle value or _____th percentile. _____% of the data are below that value.

The first quartile (Q_1) is the _____th percentile. _____% of the data are below that value.

The third quartile (Q_3) is the _____th percentile. _____% of the data are below that value.

5-Number Summary – Described by the minimum, Q_1 , median, Q_3 , and the maximum values for a data set.

Range: Describes the distance between the minimum and maximum value

$$\text{Range} = \text{Max} - \text{Min}$$

Interquartile Range or IQR (Spread): The width of the middle 50% of the data

$$\text{IQR} = Q_3 - Q_1$$

Boxplots

Example: The data set below represents 15 exam scores for a fictional MTH 243 Statistics class at PCC

31 62 65 70 76 81 82 82 87 88 89 94 95 98 100

How to draw a Boxplot: Some books call this a modified boxplot because outliers are shown.

- 1. Collect statistics:** Collect the 5 number summary and calculate the IQR
- 2. Draw the Box:** Determine the scale and draw vertical lines at the Median, Q1, and Q3. Connect these to form the box. Label your horizontal axis and include the scale.
- 3. Determine Outliers:** We use 1.5 times the interquartile range on each side of the box to determine the fences. Any data outside the fences are considered outliers. The whiskers are drawn to the nearest **data** values inside each fence.

$$\text{Upper Fence} = Q3 + 1.5 * \text{IQR}$$

$$\text{Lower Fence} = Q1 - 1.5 * \text{IQR}$$

NEVER DRAW THE FENCES!! They are not technically part of the graph, just bounds to determine outliers.

- 4. Draw the Whiskers:** Draw lines to the nearest data value inside each fence and make a short vertical bar. Label each value outside the fences (outliers) with a dot.

Example Continued: Continuing with our test score data set, write a paragraph describing the distribution. Be sure to talk about the shape, center and spread, and any unusual features (or say that there are none).

Comparing Different Groups with Side-By-Side Boxplots

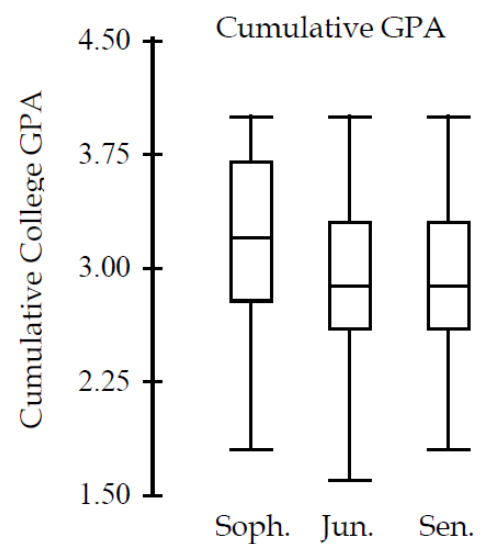
Example 4. The side-by-side boxplots show the cumulative college GPAs for sophomores, juniors and seniors taking an intro stats course.

a. Which class (sophomore, junior, or senior) had the lowest cumulative college GPA? What is the approximate value of that GPA?

b. Which class has the highest median GPA, and what is that GPA?

c. Which class has the largest range for GPA, and what is it?


d. Which class has the most symmetric set of GPAs? The most skewed set of GPAs?




Technology – Summarizing Numerical Data (Geogebra)

I will use the data representing the 15 fictional exam scores from a MTH 243 course at PCC.

31 62 65 70 76 81 82 82 87 88 89 94 95 98 100

- Visit [geogebra.org](https://www.geogebra.org)
- Under Classic Apps select Spreadsheet
- Input the numerical data into Column A
- Highlight the data and select the icon of the histogram in the upper left:  and select One Variable Analysis

Graphs

- In the upper left corner there is a drop down menu where you can select the different types of graphs: histogram, box plot, dot plot, and stem-and-leaf plot.
- For boxplots, dotplots, and stem-and-leaf plots you do not need to make any adjustments.
- For histograms you may need to adjust the starting value and the bin width.
 - Select the image of the gear at the top to open up Options 
 - Under Classes tick the box for Set Classes Manually
 - You can select the image of the gear again to close down the options
 - You should now see an option for editing “Start” which is the starting value of the graph and “Width” which adjusts the width of the bars.

Statistics

- With the graphs window open, in the upper right there is an icon for $\sum x$ which will display the statistics

n	15
Mean	80
σ	17.1542
s	17.7563
Σx	1200
Σx^2	100414
Min	31
Q1	70
Median	82
Q3	94
Max	100

Displaying a Single Categorical Variable

A frequency table uses category names for each row and records the total count of each value. A relative frequency table gives the percentage in each category.

- a. Using a group's data, create a frequency table and relative frequency table with "Eye Color" as the categorical variable.

Eye Color	Frequency (Count)	Relative Frequency (%)
Blue	5	
Brown	13	
Green	2	
Other	3	

Bar Charts**Pie Charts**

Displaying Two Categorical Variables

Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below. Source: http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Marginal Probabilities (Margins or Totals)

a. If we selected a participant at random, what is the probability that they had a relapse?

b. What is the probability that a participant did not have a relapse?

These are called marginal probabilities because we use the numbers in the margins. The use the total for a single variable.

Joint Probabilities (And)

c. What is the probability that a participant took desipramine and had a relapse?

d. What is the probability that a participant had the placebo and had a relapse?

These are called joint probabilities because they are the intersection between two variables. They are “and” probabilities.

Basic Probability Definitions

Example: Imagine we have a sample of 10 jelly beans (1 blue, 2 purple, 3 green, and 4 red). From this set of 10, we are going to draw one jelly bean.

- **Experiment** – An attempt to produce a random phenomenon where we know what outcomes could happen, but not which particular values will happen.
- **Event** – A collection of outcomes. Usually, we identify events so that we can attach probabilities to them.
- **Sample Space** – The collection of all possible outcome values. The collection of values in the sample space has a probability of 1.

Finding the Probability of Events

$$P(event) = \frac{\# \text{ of } \text{outcomes in event}}{\text{total } \# \text{ of outcomes}}$$

- **Probability** – The probability of an event is a number between 0 and 1 that reports the likelihood of that event's occurrence.
- **Theoretical Probability** – If we draw ten jelly beans (with replacement), what is the theoretical probability of a certain event?
- **Empirical Probability** – If we draw ten jelly beans (with replacement), what is the empirical probability of a certain event?
- **Law of Large Numbers** – The long-run empirical probability of repeated independent events gets closer and closer to the true theoretical frequency as the number of trials increases.
- **Independence** – The result of one event does not change the probability of the next event
- **Complements** – The probability of not some event $P(not A) = P(A^C) = 1 - P(A)$

Basic Probability Rules

Example: Imagine we have a sample of 10 jelly beans (1 blue, 2 purple, 3 green, and 4 red). From this set of 10, we are going to draw one jelly bean. While we have no way of knowing for certain what color of jelly bean we will draw, we can determine the numerical probability of drawing a specific color. We also determine the probability of NOT drawing that color. Lastly, we can determine the probability of drawing either of two given colors.

- A. What is the probability of drawing a green jelly bean?

- B. What is the probability of drawing a purple jelly bean?

- C. What is the probability of drawing a blue or red jelly bean?

- D. What is the probability of drawing a jelly bean that is purple and blue?

- E. What is the probability of drawing a jelly bean that is not green?

Complements are if we want to find the probability of not event A, $P(A^C) = 1 - P(A)$

- F. What is the probability of drawing a jelly bean that is not purple?

Addition Rule for Probability

Example: Imagine we have a sample of 10 jelly beans (1 blue, 2 purple, 3 green, and 4 red). From this set of 10, we are going to draw one jelly bean. What is the probability of drawing a jelly bean that is purple or red?

Disjoint Events cannot occur at the same time or share no common outcomes (a chip cannot be green and black at the same time). They are **mutually exclusive**.

“OR” Events

If A and B are disjoint events, $P(A \text{ or } B) = P(A) + P(B)$

Example: Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized to the right.

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

a. What is the probability that they were given desipramine or lithium?

b. What is the probability they had lithium or relapsed?

Non-disjoint Events can occur at the same time, meaning a person or item can hold more than one characteristic.

If two events A and B are non-disjoint, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Multiplication Rule for Probability

Using the wording of “and” in a question of probability could examine the probability of a sequence of events. When finding a probability for a sequence of events, we use multiplication.

Example: Imagine we have a sample of 10 jelly beans (1 blue, 2 purple, 3 green, and 4 red). From this set of 10, we are going to draw two jelly beans. If we draw the jelly beans with replacement, what is the probability of choosing a green jelly bean and then a red jelly bean?

If we draw jelly beans with replacement, what is the probability of choosing a red jelly bean and then a purple jelly bean?

Two events are called **independent** if knowing that one occurs does not change the probability that the other occurs.

For independent events, $P(A \text{ and } B) = P(A) \cdot P(B)$

Example: If we draw jelly beans without replacement, what is the probability of choosing a green jelly bean and then a red jelly bean?

If we draw jelly beans without replacement, what is the probability of choosing a purple jelly bean and then a green jelly bean?

10% Condition – When we sample less than 10% of the population, the probabilities can be assumed to be independent.

For example, say we were looking at all Americans and wanted to find the probability of selecting two Americans with blue eyes given that 17% of Americans have blue eyes and there is a population of 327.2 million.

Independence Test Example – Multiplication Rule

For independent events, $P(A \text{ and } B) = P(A) \cdot P(B)$

We can use this rule as a test. For theoretical probabilities, if the two sides of the equation are equal, the two events are independent. If the two sides are not equal, they are dependent.

With empirical data, the two sides would rarely be equal, but if they are close, they are independent. If they are significantly different, they are dependent. How close is close? We don't have the tools for that yet, so just explain your answer.

Note: this is a crude test for now to get at the concept. There is a significance test using the whole table in Math 244.

Example: Is having a relapse independent of the type of treatment?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Conditional Probability

Example: Suppose that of the 24 students in a class, 82% like chocolate, 59% like espresso, and 53% like both chocolate and espresso. What's the probability that a randomly chosen student likes chocolate given we know that they like espresso already?

Conditional Probability Formula: For events A and B,

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

For **non-independent events** $P(A \text{ and } B) = P(A) * P(B|A)$

Example. Researchers randomly assigned 72 chronic users of cocaine into three groups: desipramine (antidepressant), lithium (standard treatment for cocaine) and placebo. Results of the study are summarized below. Source: http://www.oswego.edu/~srp/stats/2_way_tbl_1.htm

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

- a. If a person took desipramine, what is the probability that they had a relapse?

$$P(\text{Relapse} | \text{Desipramine}) =$$

- b. Given that a person had lithium, what is the probability that they had a relapse? Write the probability statement and the answer.

If $P(B|A) = P(B)$, **then A and B are independent.** This means knowing that event A occurred does not affect the chance of B occurring.

Example: Is taking desipramine independent of relapsing?

Independence Test Example – Conditional Test

If $P(B | A) = P(B)$, then A and B are independent. This means knowing that event A occurred does not affect the chance of B occurring.

Just like the multiplication rule, we can use this rule as a test. For theoretical probabilities, if the two sides of the equation are equal, the two events are independent. If the two sides are not equal, they are dependent.

With empirical data, the two sides would rarely be equal, but if they are close, they are independent. If they are significantly different, they are dependent. How close is close? We don't have the tools for that yet, so just explain your answer.

Note: this is a crude test for now to get at the concept. There is a significance test using the whole table in Math 244.

Example: Is having a relapse independent of the type of treatment?

	relapse	no relapse	total
desipramine	10	14	24
lithium	18	6	24
placebo	20	4	24
total	48	24	72

Probability Practice

How are the smoking habits of students related to their parent's smoking? Here is a contingency table of data from survey of students in 8 Oregon high schools.

	Two parents smoke	One parent smokes	No parents smoke	Total
Student smokes	400	416	188	1004
Student does not smoke	1380	1823	1168	4371
Total	1780	2239	1356	5375

- $P(\text{student smokes})$
- $P(\text{no parent smokes})$
- $P(\text{at least 1 parent smokes})$
- $P(\text{student smokes and 1 parent smokes})$
- $P(\text{student smokes and no parent smoke})$
- $P(\text{student smokes \& at least 1 parent smokes})$
- What is the probability that a student who smokes has no parents that smoke?
- What is the probability that if two parents smoke, their child will smoke?
- Do parents smoking and the student smoking appear to be independent of each other?

The Binomial Model**Bernoulli Trials**

Repeated trials of an experiment are called Bernoulli trials if the following conditions are met:

1. Each trial has only two possible outcomes (generally designated as success and failure)
2. The probability of success, p , remains the same for each trial
3. *The trials are independent. (The outcome of one trial has no influence on the next)

Example A basketball player makes about 82% of her free throws. Assuming the shots are independent, find the probability that in tonight's game she will do the following:

a. Make the first free throw on the 4th attempt?

b. Make the first free throw on the 12th attempt?

In part a above, we looked at the probability that the basketball player made her first basket on the 4th try. What if we wanted to know the probability that she made exactly one basket in the first four attempts? (Note: Which of the first four doesn't matter, just that she makes one!) Write down all possible combinations that she makes one basket out of four:

Binomial Probability Model for Bernoulli Trials $X \sim \text{Binomial}(n, p)$

p = probability of success

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

X = the **number of successes in n trials**

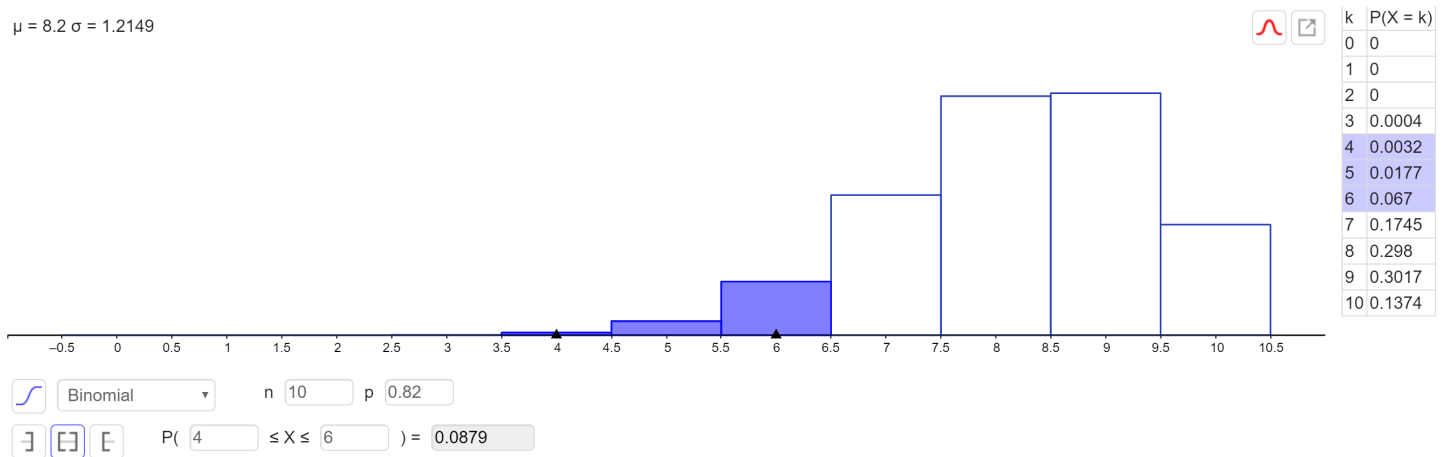
Expected Value:

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

Standard Deviation:

$$\sigma = \sqrt{npq}$$

$\mu = 8.2$ $\sigma = 1.2149$



Example Our free thrower is still making 82% of her baskets. Assume each shot is independent of the last. She's going to shoot 10 free throws.

a. What's the probability that she makes exactly 5 free throws?

b. What's the probability that she makes 9 or 10 free throws?

c. What's the probability that she makes at most 2 free throws?

d. What's the probability that she makes at least 7 free throws?

e. What's the probability that she makes between 6 and 8 free throws?

d. What's the expected number of baskets she makes? What's the standard deviation?

An Olympic Archer is able to hit the bull's-eye 80% of the time. Assume each shot is independent of the others. If he shoots 6 arrows, what's the probability of each of the following results?

- a. His first bull's-eye comes on the third arrow.
- b. He misses the bull's-eye at least once.
- c. His first bull's-eye comes on the fourth or fifth arrow.
- d. He gets exactly 4 bull's-eyes
- e. He gets at least 4 bull's-eyes.
- f. He gets at most 4 bull's-eyes
- g. How many bull's-eyes do you expect him to get? With what standard deviation?

Discrete Random Variables

We're going to study **random variables** in this chapter, which are variables whose numeric value is based on the outcome of an event.

X = a random variable that can represent each of the possible outcomes

Expected Value: $\mu = E(X) = \sum x \cdot P(x)$

Standard Deviation: $\sigma = SD(X) = \sqrt{\sum (x - \mu)^2 \cdot P(X)}$

Example: A café observed the following distribution for the number of cups of coffee sold in a day. How many cups can they expect to sell on average?

Cups Sold	30	36	42	48	60
Probability	.05	.20	.25	.15	.35

Example Find the standard deviation.

Discrete Random Variables Practice

Example You draw a card from a deck. If you get a red card, you win nothing. If you get a spade, you win \$5. For any club you win \$10, if the card is ace of clubs you get an extra \$10.

- a. Create a probability model for the amount you win.
- b. Find the expected amount you'll win.
- c. Find the standard deviation for the amount you'll win.

Practice

1. The probability model below describes the number of repair calls that an appliance repair shop may receive during an hour.

$X = \# \text{ of repair calls}$	0	1	2	3
$P(X = x)$	0.1	0.3	0.4	

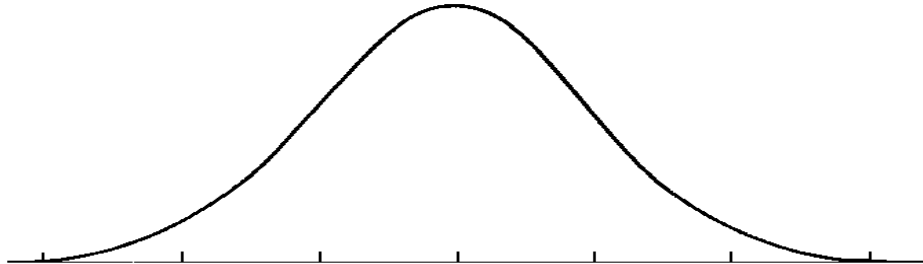
- a. Complete the table.
- b. How many calls should the shop expect per hour? Include units on all means.
- c. What is the standard deviation? Include units on all standard deviations.
- d. What is the probability of being within one standard deviation of the mean?

The Normal Model

The Normal Model – The famous bell curve

Example 1. The mean annual rainfall in Portland is unimodal and approximately symmetric with a mean of 40 inches and a standard deviation of 8 inches, rounded to the nearest inch. Label the Normal distribution model for this situation.

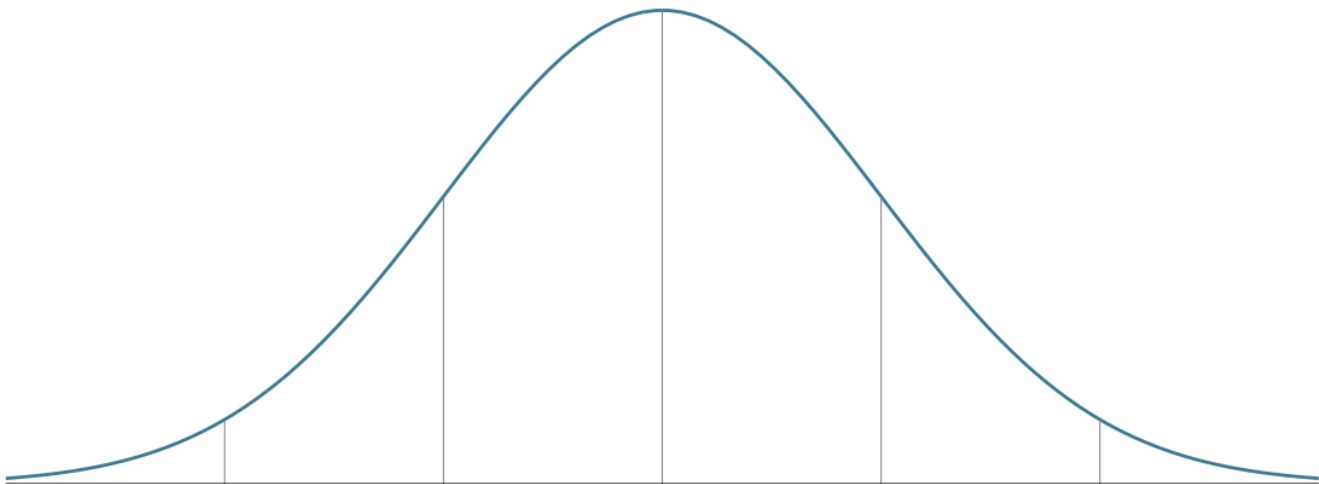
$$X \sim N(40, 8)$$



68-95-99.7% Rule

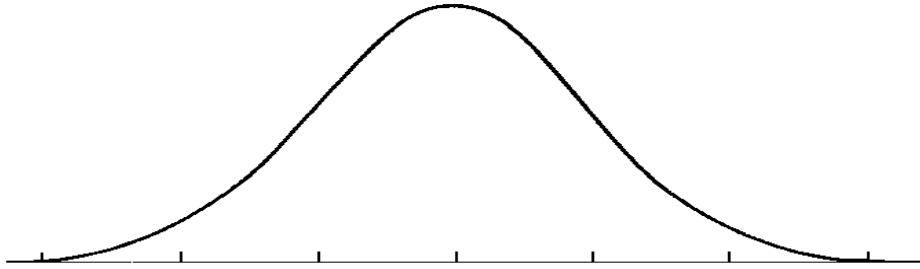
In the normal model, about 68% of the values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean. This is also called the **Empirical Rule**. Label the bell curve above to show these key features.

A value that is more than two standard deviations away from the mean is considered **unusual** or an **outlier**. A value that is more than three standard deviations away from the mean is **very rare**.



Finding Probabilities using the Empirical Rule

Example 1 continued. To find a percentage using the Normal model, draw and label the model and shade the area or percentage that you want to find. In this case our mean annual rainfall in Portland is 40 inches with a standard deviation of 8 inches.



- a. What percentage of the time is the annual rainfall between 32 and 48 inches?
- b. What is the probability that the annual rainfall is more than 48 inches?
- c. What percentage of the time is the annual rainfall 24 inches or less?
- d. What is the probability that the annual rainfall is 56 inches or less?
- e. What is the probability that the annual rainfall is 54 inches or less?

Finding Probabilities using OnlineStatbook

Go to: http://onlinestatbook.com/2/calculators/normal_dist.html

Finding Normal Probabilities

Type in the values for the mean, μ , and the standard deviation, σ . Then select above, below, between or outside depending on which probability you need to find, enter your value(s) and press enter.

Note: you can find these with GeoGebra, but it doesn't label the graph with 3 standard deviations on each side. If you copy and paste an image into a report or assignment, copy it from OnlineStatbook.

Activity: Find the exact probabilities for example 1.

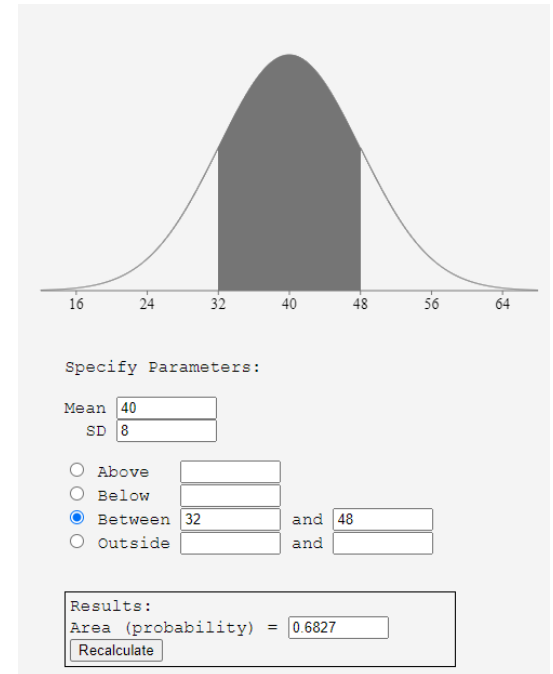
a.

b.

c.

d.

e.



Practice. In a medical study the population of children in Wisconsin were found to have serum cholesterol levels that were normally distributed with a mean $\mu = 1.75$ mg/ml and a standard deviation $\sigma = 0.30$ mg/ml.

a. Define and draw the Normal model for children's cholesterol in Wisconsin.

b. A child has a cholesterol level of 2.11 mg/ml. What is the percentage of children in Wisconsin who have cholesterol levels that are higher than this child's?

c. Find the percentage of children in Wisconsin who have cholesterol levels between 1.30 mg/ml and 2.23 mg/ml.

Percentiles and Cutoff Values

For any percentage of data, there is a corresponding **percentile** or **cutoff value**. That is the value that leaves the given percentage of data below it. We may be given a percentage and need to find the cutoff value or cut-score. Note that a **percentile** is a cutoff value, not a percentage.

Finding Cutoff values using OnlineStatbook

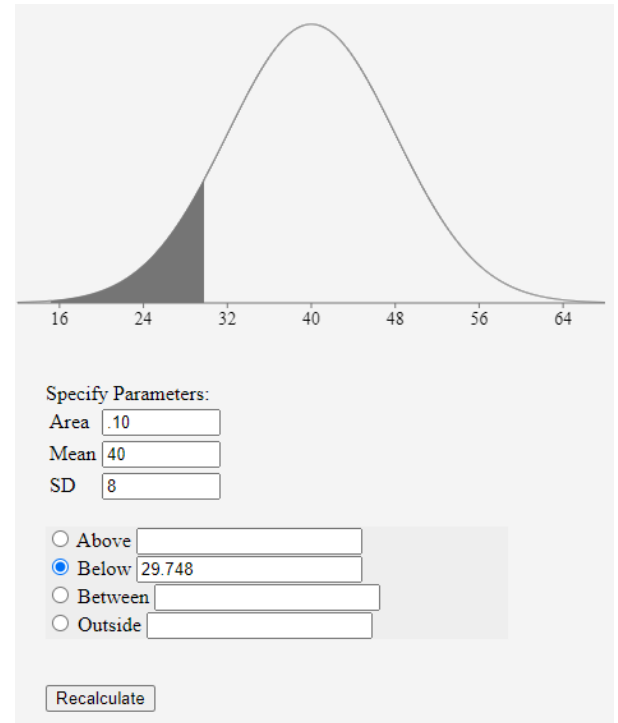
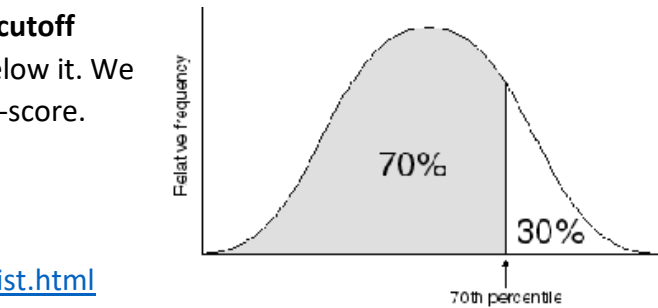
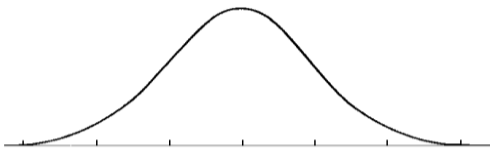
Go to: http://onlinestatbook.com/2/calculators/inverse_normal_dist.html

Type in the percentage or percentile value as a decimal (area), the mean, μ , and the standard deviation, σ . Select above, below, between or outside and press enter.

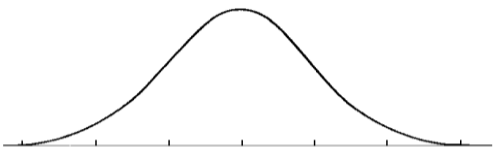
Note: you can find these with Geogebra also, but it does not label the scale correctly. If you copy and paste an image, copy it from OnlineStatbook.

Example 2. Let's continue the rainfall example where the mean annual rainfall in Portland is 40 inches with a standard deviation of 8 inches. Shade and find the cutoff values for:

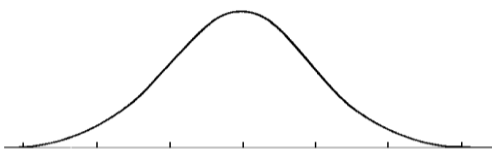
a. The lowest 10% of rainfall (the 10th percentile).



b. The highest 5% of rainfall (the 95th percentile).



c. The middle 50% of rainfall.



The Standard Normal Model, Z

Now we want to compare unlike events: Even if two events are quite different, we can still compare them using the standard deviation as a ruler. We can see how many standard deviations each event is away from its mean.

Example Assume the average annual rainfall for in Portland is 40 inches per year with a standard deviation of 8 inches. Also assume that the average wind speed in Chicago is 10 mph with a standard deviation of 2 mph. Suppose that one year Portland's annual rainfall was only 24 inches and Chicago's average wind speed was 13 mph. Which of these events was more extraordinary?

Z-score Formula: $z = \frac{x - \mu}{\sigma}$

z-score for 24 inches of rain in Portland:

z-score for a wind speed of 13 mph in Chicago:

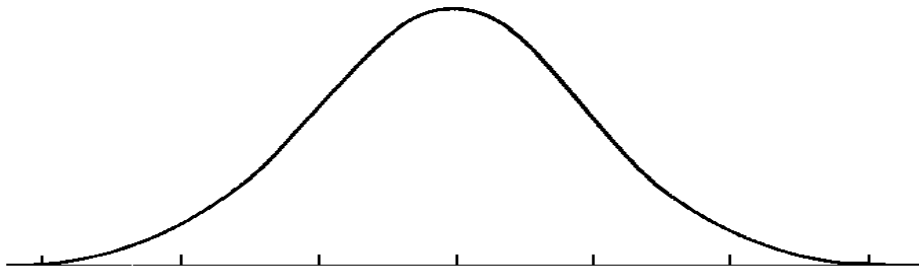
Standardizing into z-scores:

- Does not change the *shape* of the distribution of the variables
- Changes the *center* by making the mean 0.
- Changes the *spread* by making the standard deviation 1

The **Normal Distribution Model** is one that models the continuous distribution of populations. It's identified as a "bell-shaped curve." **Normal models** are only appropriate for distributions that are symmetric and unimodal. The **Standard Normal Model** has a mean of 0 and a standard deviation of 1. We denote this with $N(0,1)$. In general, a Normal model with mean μ and standard deviation σ is denoted with $N(\mu, \sigma)$. For a normal model, the z-score is given by:

$$z = \frac{x - \mu}{\sigma}$$

Standard Normal, $Z \sim N(0,1)$



Example: Find $P(z \leq 1.5)$

Example: Find $P(-.3 \leq z \leq .5)$

Example An incoming freshman took her college's placement exams in French and mathematics. In French, she scored 82 and math 86. The overall results on the French exam had a mean of 72 points and a standard deviation of 8 points, while the mean math score was 68 points, with a standard deviation of 12 points. On which exam did she do better compared with the other freshman?

Sampling Distribution of a Mean Experiment

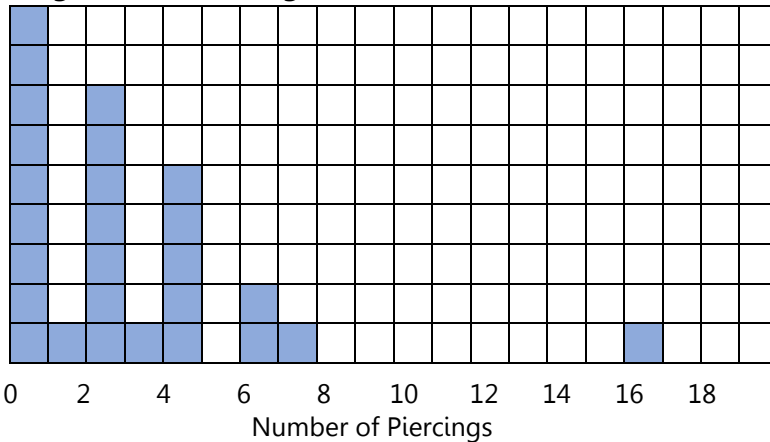
The distribution of random samples

Random samples have their own distribution models and we need to understand what they look like before we can make inferences.

This is a histogram of the number of piercings each student had in a Math 243 class. Let's take a random sample of 2 students from our class and take the average. If we take many samples of 2 students, what will the histogram look like?

Population mean, $\mu =$

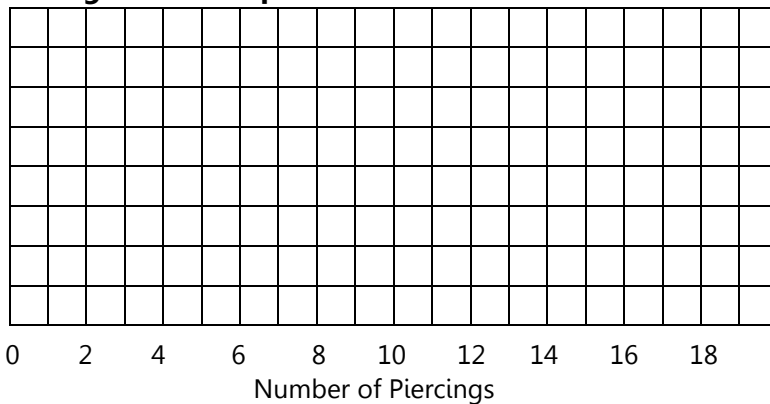
Original Class Histogram



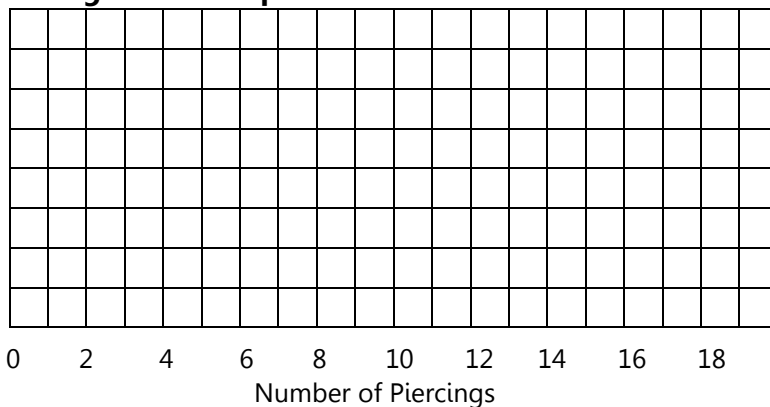
Random Sample

Sample Mean, \bar{x}

Histogram of Sample Means with n=2



Histogram of Sample Means with n=5



Sampling Distribution of Means

1. Watch as I choose a sample of size 25. What is the relationship between the various windows?

Parent Population:

Sample Data:

Distribution of Means:

2. Look at the statistics posted on the left of each window. For each one, what exactly does the “mean” refer to? Why do they differ?
3. Describe the shape of the histogram created when I click on the button to take samples 10,000 at a time. In what ways does the distribution differ significantly from the “parent” distribution at the top of the page?
4. What if the parent population is skewed? What does the distribution of means look like with $n=2$? What does the distribution of means look like with $n=25$?

Central Limit Theorem

The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.

The Central Limit Theorem requires the following conditions to hold:

- Independence Assumption: The sampled values must be independent of each other.
- Randomization Condition: The samples need to be randomly chosen, or it’s not safe to assume independence.
- Sample Size Condition: A Normal model is appropriate if a sample is **large enough**. This is very vague and is determined on a case-by-case basis. In general, the more skewed a distribution is, the larger the sample needs to be.

Properties of a Sampling Distribution of Means**The Sampling Distribution Model of Means**

When a random sample is drawn from a population with mean μ and a standard deviation, σ , we have the following mean and standard deviation for the sampling distribution:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = SD(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Regardless of the shape of the population distribution, the shape of the sampling distribution will be approximately Normal as long as the sample size is large enough.

Starting with a Population that is Normally Distributed

Example 1. A person's measured glucose level one hour after ingesting a sugary drink varies according to the Normal distribution with $\mu = 125$ mg/dl and $\sigma = 10$ mg/dl.

a. If a single glucose measurement is made, define and draw the distribution. What's the probability that a single measurement is greater than 140mg/dl?

b. If measurements are made on three different occasions and the mean result is computed, discuss each of the conditions required to use a sampling distribution for the average of the three results.

c. Define the sampling distribution model and its parameters. Draw and label the model relative to the model for part a.

d. What's the probability that the mean of three measurements is greater than or equal to 140mg/dl?

e. What is the 95th percentile for the average of three results from this person?

Properties of a Sampling Distribution of Means continued**Starting with a Population that is Not Normally Distributed**

Example 2. Restaurant bills at a given restaurant have an assumed population mean of \$32.40 and a population standard deviation of \$8.16. This data is heavily skewed to the left.

- a. Explain why you cannot determine that a given bill will be at least \$35.

- b. Can you estimate the probability that the next 5 bills will average at least \$35? Discuss each of the conditions for using the sampling distribution of the mean.

- c. If we take the average of the next 50 bills, would all the conditions be met?

- d. Define the model with its parameters. Draw and label it.

- e. How likely is it that the next 50 bills have an average of at least \$35?

- f. Find the two values for the middle 50% of the average of 50 bills.

The number of piercings was **quantitative data** so we found a sampling distribution for the mean of each sample. Now we are going to make a sampling distribution for **categorical (yes/no) data**. We need to see what random samples of proportions look like before we can make inferences. Let's take an anonymous poll of our class. If we take samples of 3 students, what will the histogram look like? What if we take samples of 5 students?

Statistical question: _____

Sample Proportion, \hat{p}

Population proportion, $p =$ [illegible]

Proportion in the sample who _____

[illegible]

Proportion in the sample who _____

Sampling Distribution of Proportions

Summarize the simulation results for drawing random samples from a population where 62.5% of the population are born in Oregon.

Original Population:

Sample:

Distribution of Sample Proportions:

1. What (in general) happens to the sample proportion as the sample size gets larger?
2. What (in general) happens to the sample standard deviation as the sample size gets larger?
3. For proportions, there is not a single minimum sample size number like we have for means. What happens when you start with a sample proportion of 0.5? Try 0.1.

Sampling Distribution of Proportions Properties

The Sampling Distribution Model for a Proportion

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution \hat{p} is modeled by a Normal model with mean p and standard deviation.

$$\mu_{\hat{p}} = p \qquad \sigma_{\hat{p}} = SD(\hat{p}) = \sqrt{\frac{pq}{n}}$$

The Normal model is an appropriate approximation for a sampling distribution of proportions if the following conditions hold:

- Independence Assumption: The individuals in the sample (whose proportions we are finding) must be independent of each other. If we sample less than 10% of the population, we can assume the events are independent.
- Randomization Condition: The samples need to be randomly chosen, or it's not safe to assume independence.
- Success/Failure Condition: You should have at least 10 successes and 10 failures in your data

$$n \cdot p \geq 10$$

$$n \cdot q \geq 10$$

Example: Suppose 60% of all college students are either very satisfied or completely satisfied with their online shopping experience. You decide to take a nationwide random sample of 2500 college students and ask if they agree. Is it reasonable to use a Normal model for the sampling distribution of the sample proportion?

- Find the mean and standard deviation of the proportion of survey respondents who agree
- What's the probability that the sample proportion who agree is at least 57%?

Information on a packet of seeds claims that the germination rate is 92%. There are approximately 160 seeds in each packet.

- Discuss each of the conditions required to use a sampling distribution for the proportion of seeds that will germinate.
- Define the sampling distribution model and its parameters.
- Sketch and label the model.
- What is the probability that more than 95% of the seeds will germinate?
- Would it be unusual for only 140 or fewer of the 160 seeds to germinate? Explain.

Introduction to Confidence Intervals for Proportions

In this section we'll look at what a sample proportion (\hat{p}) tells us about the true population proportion (p). We won't be able to know anything for sure but we can find a range of values that are *reasonable* for our population value.

The **standard error** is an estimate of the standard deviation of the sampling distribution of a proportion ($SD(\hat{p})$). It's used when we don't know the value of p and are not able to determine the true standard deviation.

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

Example 1

A survey of 2,500 HR professionals showed that 1,200 use social networking sites to research job applicants. State the sample proportion and calculate the standard error for this sample proportion.

Draw how the sampling distribution model for \hat{p} should look.

Based on the 68-95-99.7 Rule, where do you expect 95% of sampling distributions \hat{p} to fall.

Example 2

Between what two proportions do we capture **exactly** 95% of values. Write the interval that represents the proportions.

Confidence Interval for Proportions

sample proportion \pm margin of error

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

This creates an interval of possible values for the true value of the population proportion

The **critical z-score** is denoted by z^* and is determined by the confidence level.

The **margin of error** is given by $ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Wording Conclusions and Finding Critical Z-Scores**Example 3**

Can you say the following (or not)?

- a. 48% of all hr professionals us social networking sites to research job applicants
- b. It is probably true that 48% of all hr professionals use social networking sites to research job applicants
- c. We don't know exactly what proportion of all hr professionals use social networking sites to research job applicants, but we know that it's between 46% and 50%
- d. We don't know exactly what proportion of all hr professionals use social networking sites to research job applicants, but it's probably between 46% and 50%
- e. We are 95% confident that between 46% and 50% of all hr professionals use social networking sites to research job applicants.

Example 4

Identify the critical z-scores for 99% confidence, 95% confidence, 90% confidence, and 80% confidence. (These will be helpful throughout this chapter!)

Example 5

a. Construct a 99% confidence interval for the true proportion of US adults that are baseball fans and interpret.

- b. Construct and 80% confidence interval for the true proportion of US adults that are baseball fans and interpret.

What happens to the size of the confidence interval the more certain you are (the higher the confidence level)?

What happens to the size of the confidence interval the less certain that you are (the lower the confidence level)?

$$n = \hat{p} \cdot \hat{q} \cdot \left(\frac{z^*}{ME} \right)^2$$

Example 6

b. Suppose we want to cut the margin of error to 4% (again with 90% confidence). What's the necessary sample size?

Given we build a distribution for samples with some assumed value for p , does our \hat{p} provide evidence the distribution is false?

During the 2013 NFL season, the home team won 153 of 245 regular-season games. In this chapter, we'll determine if this is evidence of a home-field advantage. As you can probably guess, "evidence" is a bit ambiguous. What we'll do is determine the probability that this is due to natural sampling variation *given the assumption that there was NO home field advantage*.

a. Write the null hypothesis and alternative hypothesis for the NFL home field advantage problem.

b. To decide if there is a home field advantage, what probability do we need to determine?

c. Calculate the above probability. Include a picture in your work!

In this example, we determined the probability that our observed sample statistic occurred given our null hypothesis is true. This is known as the **P-value**

At this point we have two choices:

1. Reject the null hypothesis. We do this when the p-value is *small*
 2. Fail to reject the null hypothesis. We do this when the p-value is not sufficiently *small*
-
- d. Determine if you should reject the null hypothesis or fail to reject the null hypothesis for the home field advantage problem. Write a clear summary.

Note: Since we are using the Normal model, we need our four conditions to be satisfied: Independence, Randomization, the 10% Condition, and the Success/Failure Condition.

Hypothesis Testing: One-Proportion z-Test

1. State the null and alternative hypotheses
2. Check the conditions required to use a Normal model
3. Calculate the test statistic & p-value to evaluate
4. Write a summary.

Example 3: For each scenario below, state the null hypothesis and the alternative hypothesis

- a. An old antacid provided relief for 70% of the people who used it. A pharmaceutical company has a new drug they want to test to see if it is more effective.
- b. A company is concerned that too few of its cars meet pollution emission standards. They have 1500 cars. They sample 56 cars and find that 43 meet emissions standards. They want to test if less than 80% of their fleet meets emissions standards.
- c. In 2014, the official poverty rate was 14.8% in the US. A city official wants to test if their county has a different poverty rate than the rest of the US.

So far, we've been a bit vague with regard to a p-value being *small enough*. This is because how small a p-value needs to be depends on the context. The **significance level** (α) is the threshold for which we can reject a null hypothesis. The most common significance level is $\alpha = 0.05$. This means we reject the null hypothesis as long as the p-value is *below* 0.05. You may see larger significance levels such as $\alpha = 0.10$ when the results don't have a very serious effect. In medicine (where the consequences of missing something can be dire), you will see smaller significance levels, such as $\alpha = 0.01$.

Example 4: For each scenario, decide whether to reject the null hypothesis or fail to reject the null hypothesis based on the stated p-value. Include an appropriate conclusion.

- a. A company is concerned that too few of its cars meet pollution emission standards. They have 1500 cars. They sample 56 cars and find that 43 meet emissions standards. They want to test if less than 80% of their fleet meets emissions standards. They run a hypothesis test and find that the p-value is 0.0009.
- b. In 2014, the official poverty rate was 14.8% in the US. A city official wants to test if their country has a different poverty rate than the rest of the US. He runs a hypothesis test and determines the p-value to be 0.273.
- c. Drugs that prevent seizures sometimes have unwanted side effects, such as depression. A researcher wants to study such a drug and look for evidence that such side effects exist. She runs a hypothesis test and finds a p-value of .06.

STEPS FOR HYPOTHESIS TEST FOR A SINGLE PROPORTION

STEP 1: STATE THE NULL AND ALTERNATIVE HYPOTHESES

STEP 2: CHECK THE CONDITIONS & SET UP

STEP 3: FIND THE TEST STATISTIC & P-VALUE TO EVALUATE

STEP 4: CONCLUSION

Example 5: A current antacid provides relief for 70% of the people who use it. A pharmaceutical company has a new drug and they want to test whether it is more effective. They run a study with 100 randomly selected patients and 75 people experienced relief. Is this evidence that the new drug is better?

Step 1: Write the hypotheses

Step 2: Check the conditions to use a Normal distribution

Step 3: Calculate the p-value

Step 4: Conclusion

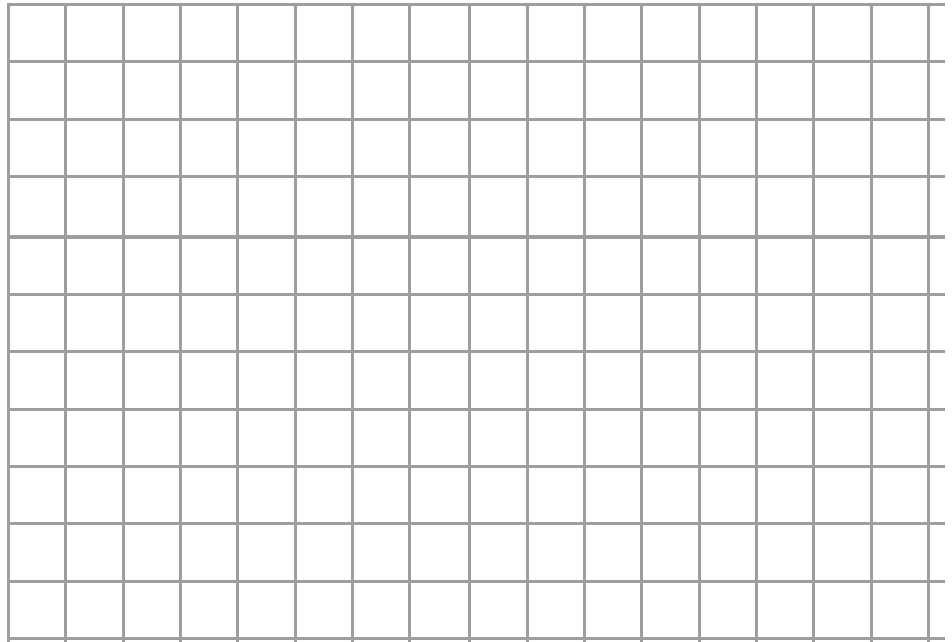
Example 6: A company develops what it hopes will be better instructions for its customers to set up their smartphones. The goal is to have 96% of customers succeed. The company tests the new system on 400 people, of whom 376 were successful. Is this strong evidence that the new system fails to meet the company goal?

Linear Regression

A **scatterplot** is a plot of paired (x, y) quantitative data with a horizontal x-axis and vertical y-axis. The pattern of the plotted points is often helpful in determining whether there is a relationship between the two variables.

Example 1: Create a scatterplot that involves the number of cricket chirps per second paired with temperatures ($^{\circ}\text{F}$).

Temperature	Chirps per Second
89	20
93	20
81	17
70	15
69	15
80	15
76	14
72	16
84	19
75	16
82	17
83	16
83	17
84	17
81	16



- Is an increased number of chirps per second **caused** by an increase in temperature?
- What happens (in general) to the number of chirps per seconds as temperature increases?

Explanatory and Response Variables

The **response variable** is the dependent variable (y). In our example:

The **explanatory variable** is the independent variable (x). In our example:

We think that changes in the explanatory variable might explain changes in the response variable.

Caution: Even if it appears that y can be "predicted" from x , it does not follow that x **causes** y .



Example 2: There are four main features we want to analyze when looking at a scatterplot

- **Direction**: positive, negative, neither
- **Form**: straight (aka linear), curved, or no pattern
- **Strength**: strong, moderate, weak
- **Outliers**: exists or does not or groupings

Correlation

A **correlation** exists between two variables when the values of one variable are somehow *associated* with the values of the other variable.

The **linear correlation coefficient (r)** measures the strength of the linear correlation between the paired quantitative x- and y-values in a sample.

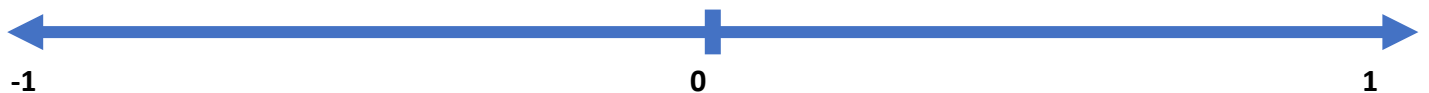
The formula:

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y} \text{ or } r = \frac{\Sigma z_x z_y}{n-1}$$

But we are going to use technology...YAY!

Characteristics of r

- The value of r is between -1 and 1.
- r has no units
- The closer r is to -1 or 1, the stronger the association.
- The sign of r is the direction of the association.



Back to Example 1:

- a. What is the correlation between the number of chirps per second with crickets and weather? (don't forget to check the conditions!)
- b. Write a brief description of the association
- c. Do these results confirm that the increase in temperature leads to a higher number of chirps from crickets? Explain.

The Coefficient of Determination, r^2 or R^2

The coefficient of determination tells us the percentage of the variation in the y-values that can be explained by the least squares regression line of y on x.

Going back to our example from last class the correlation between the number of chirps per second with crickets and temperature was $r = .8135$. Find the coefficient of determination.

Example 3: The data below shows the cost of the airfare and the distance traveled to each destination from Baltimore, MD.

<u>Destination</u>	<u>Distance</u>	<u>Airfare</u>
Atlanta	576	178
Boston	370	138
Chicago	612	94
Dallas	1216	278
Detroit	409	158
Denver	1502	258
Miami	946	350
New Orleans	998	188
New York	189	98
Orlando	787	179
Pittsburgh	210	138
St. Louis	737	98

a. Which is the explanatory and which is the response variable?

b. Create a scatterplot for the data using technology. Do we pass the conditions?

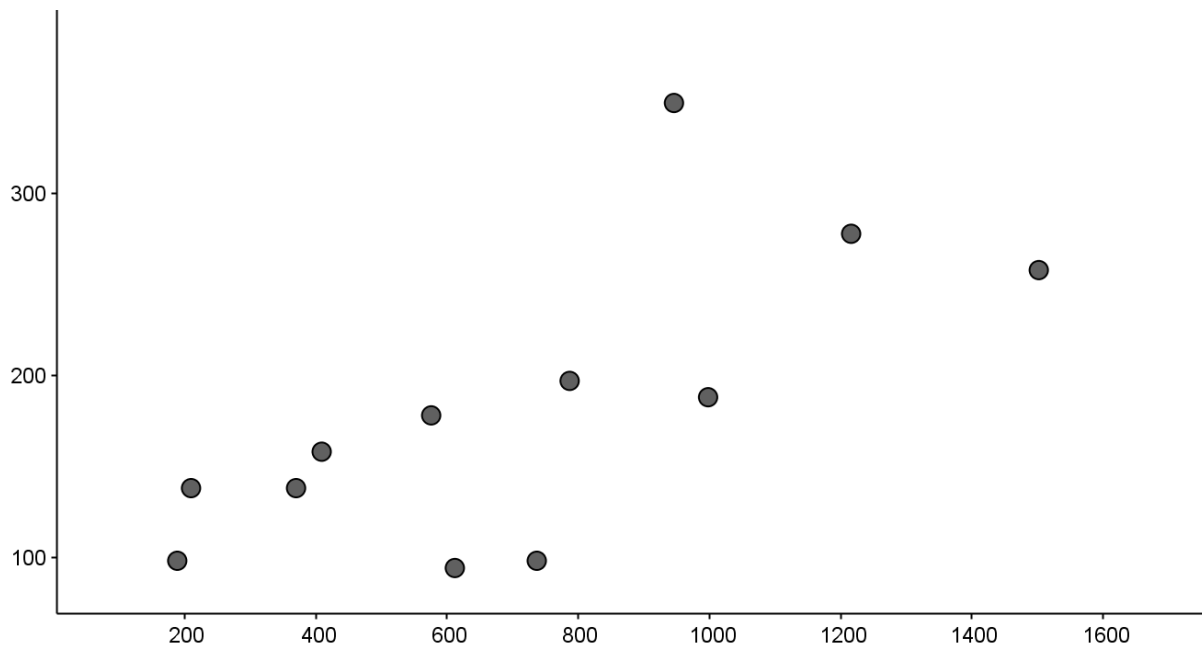
c. What is the correlation between the airfare and the distance of the flight?

d. Write a brief description of the association

e. What is the coefficient of determination?

f. Interpret the coefficient of determination.

Here is the data from the last example



The Line of Best Fit

Use a ruler or straightedge to draw a line that models this relationship

Draw the vertical distance between each point and the line. These are the residuals

$$\text{Residual} = \text{Observed Value} - \text{Predicted Value}$$

The least squares regression line is the line that minimizes the residuals.

Least Squares Regression Line - $\hat{y} = b_1x + b_0$

\hat{y} is the predicted y-value for a given value of x

b_1 is the slope of the regression line. It is the expected increase/decrease on average of the y-variable for a one-unit increase in the x-variable.

b_0 is the y-intercept of the regression line. It is the y-value when $x=0$. Sometimes it does not make sense in context.

Back to Example 3

- a. Find the least squares regression line for the data

- b. Give the slope with units and interpret in context.

- c. Give the y-intercept and interpret it in context.

- d. Estimate the cost of a ticket if they are travelling 200 miles from Baltimore, MD.

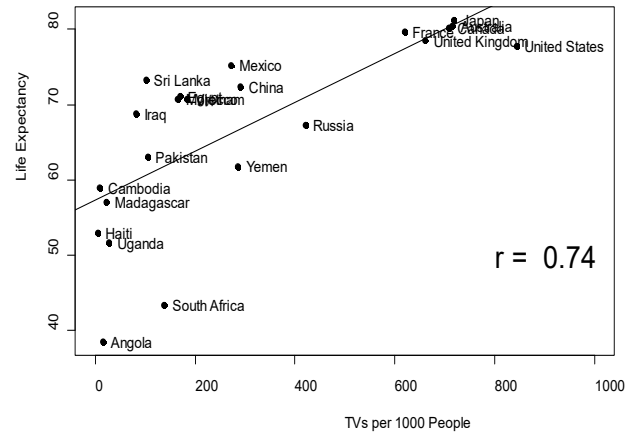
- e. Which airfare has the largest residual? What is the value of the residual?

Lurking Variables

A lurking variable is a hidden variable that is responsible for the apparent association.

A study of mortality rate across countries world wide noticed an association between average life span and the average number of televisions owned.

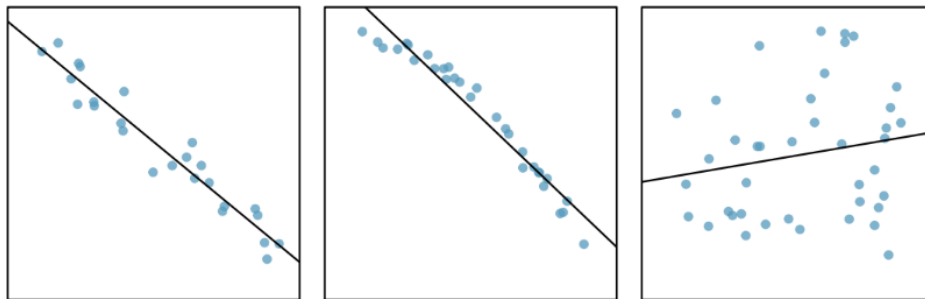
What are some possible lurking variables?



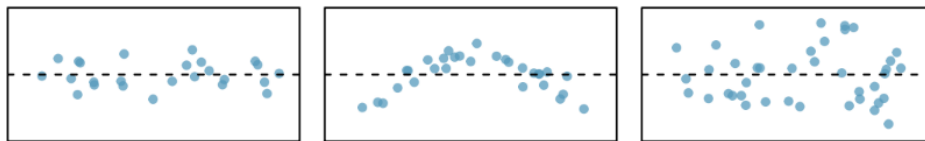
Residual Plots: $(x_i, y_i - \hat{y}_i)$

The residual is the vertical distance (error) between the data point and the predicted value (point on the regression line). It is important to examine the residual plot to see if there are any patterns that we could not see by looking at the scatterplot.

Scatterplot:



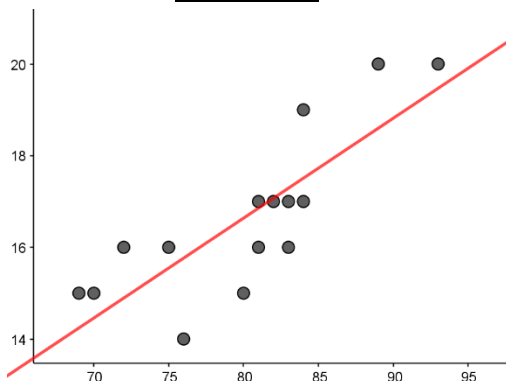
Residual:



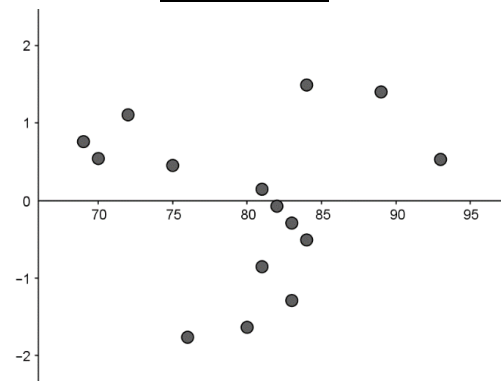
Check the residual plot for:

- Bends or other patterns in the residuals
- Outliers that weren't clear before
- Any change in spread of the residuals from one part of the plot to another

Scatterplot



Residual Plot



Making Predictions

- a. If you wanted to fly to a destination that was 500 miles from Baltimore, how much would the ticket cost according to the model?

- b. If you wanted to fly to Sydney, Australia from Baltimore (9,782 miles), do you think this model would make a good prediction?

- c. If you wanted to fly to a destination that was 10 miles from Baltimore, do you think this model would make a good prediction?

Interpolation vs Extrapolation

Interpolation is making a prediction within the x-values of your data set.

Extrapolation is making a prediction beyond the data set – be careful!! How do you know that the trend will continue?

Outlier Analysis – Let's Look at Miami

Last time we talked about Miami being a possible outlier, let's explore it again.

- a. Can you think of any reasons to explain the cost of airfare to Miami?

- b. Write down the regression equation and correlation for the data set including Miami.

- c. Perform the regression again without Miami and write the down the linear regression equation and r .

- d. Is there a meaningful difference between the two models?

- e. Should we describe Miami as an outlier?

Answer each of the following questions completely. When possible to answer using a complete sentence and offering explanation, please do so. This assignment is worth 15 points.

- Page 1

- c. Send a survey home with every student, and have parents fill it out and return it. Have teachers track the submissions and follow up with parents until all are returned.

- d. Randomly select 20 parents from each elementary school. Send them a survey and follow up with a phone call if they do not return the survey within a week.

- 3. [5 pts] Suppose PCC placed a poll online in MyPCC asking about campus climate and whether students experience discrimination. Let's say 2,952 people participated in the poll.
 - a. Describe the sample and sampling method.

 - b. Could we generalize the results of the poll to the population of all PCC students? Explain why or why not.

 - c. Explain how you would select a representative sample of PCC students that would give us results that we can generalize to the population. Give a specific identifier, method and procedure (there are many good answers).

GRADED PROBLEM SET Module 2

Answer each of the following questions completely. When possible to answer using a complete sentence and offering explanation, please do so. This assignment is worth 15 points.

Be sure to add units to all statistics and label the axes and units on all graphs ☺.

This data along the right is from the U.S. Department of Health and Human Services, National Center for Health Statistics, Third National Health and Nutrition Exam Survey. Recorded are the pulse rates in beats per minute (bpm) from people who identify as female.

a. [3 pts] Fill in the table below outlining the summary statistics, including units.

Sample Size	
Mean	
Standard Deviation	
Minimum	
First Quartile	
Median	
Third Quartile	
Maximum	

b. [2 pts] Construct a histogram for the data starting at 60 and using a bin width of 10, either by hand or using technology. Give a title and units.

c. [1 pts] Interpret the standard deviation in a complete sentence using the numerical value, units and the context.

60
60
60
64
64
64
64
68
68
68
68
72
72
72
72
72
72
72
76
76
76
76
80
80
80
80
80
80
88
88
88
88
96
104
124

- d. [2 pts] Are there any outliers in this data set? (Show your fence calculations.)
- e. [2 pts] Construct a boxplot for the data, either by hand or using technology. Give a title and units.
- f. [1 pts] Is the shape symmetric, left-skewed, or right skewed?
- g. [1 pts] Would you expect the mean or median to be larger based on the shape of the distribution? Explain.
- h. [3 pts] Summarize your findings in a paragraph by describing the shape, center, spread, and unusual features of the data, including units and context.

GRADED PROBLEM SET Module 3

Answer each of the following questions completely. When possible to answer using a complete sentence and offering explanation, please do so. This assignment is worth 15 points.

1. [8 pts] Helmets and Injuries: The data in the accompanying table is based on data from “Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders,” by Sullheim, et al. Journal of the American Medical Association, Vol. 295, No. 8.

	Head Injuries	Not Injured	Total
Wore Helmet	96	656	
No Helmet	480	2330	
Total			

- What is the probability a randomly selected person has a head injury?
- What is the probability a randomly selected person had a head injury or wore a helmet?
- What is the probability a randomly selected person did not wear a helmet and was not injured?
- What is the probability that a randomly selected person had a head injury given they did not wear a helmet?
- Are the events for helmet use and head injury independent? Prove mathematically and explain your answer.

- Note: when you use GeoGebra, write down what you are entering to show your work. The way we define a binomial variable is $X \sim \text{Binomial}(n, p)$, and fill in the numbers you entered for n and p . That way I can give you partial credit accordingly if something goes wrong 😊*

- Page 2

GRADED PROBLEM SET Module 5

Answer each of the following questions completely. This assignment is worth 15 points.

1. [4 pts] Based on past results found in the *Information Please Almanac* there is a 0.1919 probability that a baseball World Series will last four games, a 0.2121 probability that it will last five games, a 0.2223 probability that it will last six games, and a 0.3737 probability that it will last seven games.
 - a. Does the given information describe a probability distribution? Explain.
 - b. Assuming that the given information describes a probability distribution, find the mean and standard deviation for the number of games in World Series, including units.

2. [5 pts] Tree diameters in a plot of land are normally distributed with a mean of 14 inches and a standard deviation of 3.2 inches. Define, draw a label a Normal curve and find the answers using Onlinestatbook. (You only need to fully label the first drawing)
 - a. What is the probability that an individual tree has a diameter between 13 inches and 16.3 inches?
 - b. What is the cutoff score for the 95th percentile of tree diameters?

3. [6 pts] A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.
- What is the probability that a randomly chosen light bulb lasts more than 9,400 hours? Define, draw and label the distribution and give your answer in a complete sentence.
 - Let's say the distribution of the bulb lifespans is instead heavily skewed to the right. We want to select 40 bulbs and calculate their average lifespan. Write about each of the conditions needed to use the sampling distribution of a mean.
 - What is the probability that the mean lifespan of 40 randomly chosen light bulbs is more than 9,400 hours? Define, draw and label the distribution and give your answer in a complete sentence.

GRADED PROBLEM SET Module 6

Answer each of the following questions completely. All work and steps should be shown. When possible, answer using a complete sentence and offering explanation. This assignment is worth 15 points.

1. The Pew Research Center estimates that as of January 2014, 89% of 18-29-year-olds in the United States use social networking sites.
 - a. [2 pts] For a sample size of 100, write about each of the conditions needed to the sampling distribution of a proportion.

 - b. [3 pts] Define, draw and label the sampling distribution. Calculate the probability that at least 91% of 100 randomly sampled 18-29-year-olds use social networking sites and give your answer in a complete sentence.

2. ACT, Inc. reported that 74% of 1644 randomly selected college freshmen returned to college the next year.
 - a. [2 pts] Construct and interpret a 90% confidence interval to estimate the retention rate for college freshmen.

- b. [2 pts] Let's say they only need a 2% margin of error with the same level of confidence the following year. How many college freshmen must be surveyed?
2. [6 pts] There were 2430 Major League Baseball games played in 2009, and the home team won the game in 53% of the games. If we consider the games played in 2009 as a sample of all MLB games, test to see if there is evidence, at the 5% level, that the home team wins more than half the games. (Show all steps of the hypothesis test, including testing conditions and a Normal graph.)

GRADED PROBLEM SET Module 7

Answer each of the following questions completely. When possible to answer using a complete sentence and offering explanation, please do so. This assignment is worth 15 points.

1. The Nielsen Company measured connection speeds on home computers in seven different countries. The table shows the percent of internet users with a “fast” connection and the average amount of time spent online in hours per week.

Country	% Fast Connection (x)	Hours Online (y)
United States	70	26
Germany	72	28
Australia	64	23
United Kingdom	75	28
France	70	27
Spain	69	27
Italy	64	24

- a. [3 pts] Make a scatterplot and describe the association between % Fast Connection and Hours Online. Include all 4 characteristics and the correlation coefficient.
- b. [2 pts] Find the least squares regression line.
- c. [2 pts] Estimate the expected average number of hours spent online per week if 70% of homes have a fast connection.

d. [2 pts] Find the residual for the United States. Explain what it means.

e. [2 pts] Describe the slope in context with units.

f. [2 pts] Describe the y-intercept in context with units. Is this point meaningful in this context?

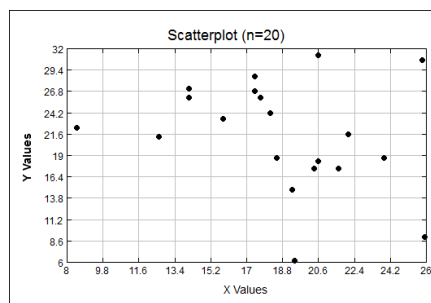
2. [2 pts] Match each scatterplot to its correlation coefficient.

A. 0.64

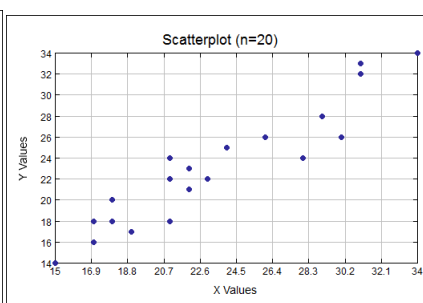
B. -0.03

C. 0.91

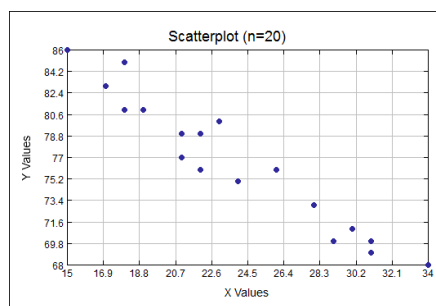
D. -0.89



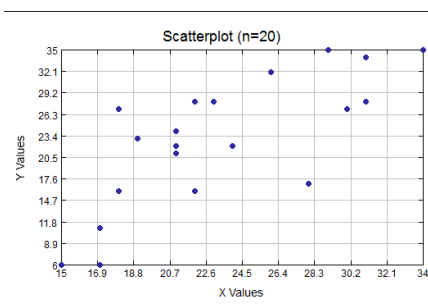
Letter _____



Letter _____



Letter _____



Letter _____